

# FREIBURGER INTERVALL-BERICHTE



80/5

Jürgen Garloff :

Untersuchungen zur  
Intervallinterpolation

1980

Institut für Angewandte Mathematik  
Universität Freiburg i. Br.



# FREIBURGER INTERVALL-BERICHTE

80/5

Jürgen Garloff:

Untersuchungen zur  
Intervallinterpolation

(Dr.-Dissertation)

Herausgeber: Karl Nickel  
Institut für Angewandte Mathematik  
Universität Freiburg i. Br.  
Hermann-Herder-Straße 10  
D-7800 Freiburg i. Br.  
West Germany  
Telefon (0761) 203 3062



Digitized by the Internet Archive  
in 2026

I n a u g u r a l - D i s s e r t a t i o n

zur

Erlangung der Doktorwürde

der

Mathematischen Fakultät

der

Albert-Ludwigs-Universität zu Freiburg i.Br.

vorgelegt von

Jürgen Garloff

aus Köln

1980

Referent: Professor Dr. K.Nickel

Korreferent: Professor Dr. J.Rokne

Tag der Verkündung des Prüfungsergebnisses: 4. 12. 1980

Debat: Prof. Dr. O. K. Kegel

Mein Dank gilt Herrn Professor Dr. Karl Nickel. Seine Anregung zum Thema, seine Hilfsbereitschaft und sein stetes Interesse haben wesentlich zum Zustandekommen dieser Arbeit beigetragen.

Danken möchte ich auch Herrn Professor Dr. Jon Rokne für die freundliche Übernahme des Korreferats.

Mein Dank gebührt auch Frau Margot Großmann, die die mühevollen Schreibarbeit übernahm, und Frau Helga Sturm, die die Zeichnungen anfertigte, sowie den Herren Dipl.-Math. Manfred Raith und Karl Reichmann für das Korrekturlesen des maschinengeschriebenen Manuskripts.



## Inhaltsverzeichnis

Seite

|           |                                                                                                                                                                                                               |    |
|-----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Kapitel 1 | Einleitung                                                                                                                                                                                                    | 1  |
| 1.1       | Aufgabenstellung und Übersicht                                                                                                                                                                                | 1  |
| 1.1.1     | Reelle Interpolation und Intervallinterpolation                                                                                                                                                               | 1  |
| 1.1.2     | Behandlung der Intervallinterpolation in der Literatur                                                                                                                                                        | 5  |
| 1.1.3     | Übersicht                                                                                                                                                                                                     | 6  |
| 1.1.4     | Bemerkungen                                                                                                                                                                                                   | 9  |
| 1.2       | Grundlegende Bezeichnungen, Definitionen und Hilfsmittel                                                                                                                                                      | 10 |
| Kapitel 2 | Optimalität der Einschließung                                                                                                                                                                                 | 20 |
| 2.1       | Optimale Funktionsstreifeneinschließung                                                                                                                                                                       | 20 |
| 2.2       | Optimale Intervalleinschließung der Parameter der Lösungsfamilie                                                                                                                                              | 23 |
| Kapitel 3 | Intervallinterpolation mit trigonometrischen Polynomen                                                                                                                                                        | 27 |
| 3.1       | Vorbemerkungen                                                                                                                                                                                                | 27 |
| 3.1.1     | Die Lebesgue-Konstante                                                                                                                                                                                        | 30 |
| 3.1.2     | Die diskrete Fouriertransformation                                                                                                                                                                            | 31 |
| 3.2       | Die Rechteck- und die Kreisscheibenarithmetik und Einschließungssätze für Produkte komplexer Intervallmatrizen                                                                                                | 32 |
| 3.2.1     | Rechteck- und Kreisscheibenarithmetik                                                                                                                                                                         | 33 |
| 3.2.2     | Der Raum der komplexen Intervallmatrizen                                                                                                                                                                      | 37 |
| 3.3       | Die intervallmäßige Durchführung der schnellen Fouriertransformation mit der Rechteck- und der Kreisscheibenarithmetik                                                                                        | 39 |
| 3.4       | Akkumulation der Rundungsfehler bei der intervallmäßigen Durchführung der schnellen Fouriertransformation und bei Bestimmung von Schranken für die Koeffizienten des trigonometrischen Interpolationspolynoms | 57 |

|           |                                                                                                                |     |
|-----------|----------------------------------------------------------------------------------------------------------------|-----|
| 3.4.1     | Maschinenintervallarithmetik                                                                                   | 59  |
| 3.4.2     | Direkte Berechnung                                                                                             | 63  |
| 3.4.3     | Schnelle Fouriertransformation mit der Rechteckarithmetik                                                      | 65  |
| 3.4.4     | Schnelle Fouriertransformation mit der Kreisscheibenarithmetik                                                 | 71  |
| 3.4.5     | Abschätzung der Konstanten $\delta_n$                                                                          | 78  |
| 3.4.6     | Numerische Ergebnisse                                                                                          | 80  |
| Kapitel 4 | Intervallinterpolation mit Splines                                                                             | 89  |
| 4.1       | Vorbemerkungen                                                                                                 | 89  |
| 4.1.1     | Die Durchführbarkeit des Intervall-Gauß-Algorithmus für einige Klassen von Intervallmatrizen                   | 91  |
| 4.2       | Kubische Splines                                                                                               | 96  |
| 4.3       | Bestimmung von Schranken für die B-Spline-Koeffizienten einer interpolierenden Spline-Funktion                 | 100 |
| 4.3.1     | Sonderfall: Periodische Splines mit äquidistanten Knoten                                                       | 107 |
| 4.4       | Totalnichtnegative Intervallmatrizen                                                                           | 112 |
| 4.4.1     | Definitionen, Bezeichnungen und Hilfsmittel                                                                    | 112 |
| 4.4.2     | Kriterien für die Totalpositivität bzw. -nichtnegativität einer Intervallmatrix                                | 116 |
| 4.4.3     | Anwendung auf lineare Intervallgleichungssysteme                                                               | 125 |
| Kapitel 5 | Intervallinterpolation mit den Funktionen $\alpha \xi^B$ und $\alpha_1 \exp(B_1 \xi) + \alpha_2 \exp(B_2 \xi)$ | 132 |
| 5.1       | Intervallinterpolation mit den Funktionen $\alpha \xi^B$                                                       | 134 |
| 5.2       | Intervallinterpolation mit den Funktionen $\alpha_1 \exp(B_1 \xi) + \alpha_2 \exp(B_2 \xi)$                    | 158 |
| Anhang:   | Triplex-ALGOL 60-Programme                                                                                     | 165 |
|           | Literaturverzeichnis                                                                                           | 173 |

## Kapitel 1 Einleitung

### 1.1 Aufgabenstellung und Übersicht

#### 1.1.1 Reelle Interpolation und Intervallinterpolation

In diesem Unterabschnitt soll zunächst dargelegt werden, was wir unter einem reellen Interpolationsproblem verstehen. Anschließend werden wir anhand eines Problems aus der Praxis die Beschäftigung mit der Intervallinterpolation motivieren und dann zwischen Intervallinterpolationsproblemen vom Typ I und vom Typ II unterscheiden.

#### Reelles Interpolationsproblem

Es sei  $D$  eine Teilmenge der Menge  $\mathbb{R}$  der reellen Zahlen und  $\mathfrak{J}$  eine Klasse von auf  $D$  definierten reellwertigen Funktionen  $\varphi$ .

Gegeben seien Stützstellen  $\xi_k \in D$  und Stützwerte  $\eta_k \in \mathbb{R}$ ,  $k = 1(1)n$ , wobei die Stützstellen als paarweise verschiedenen vorausgesetzt seien.

Gesucht ist eine Funktion  $\varphi \in \mathfrak{J}$  mit

$$\varphi(\xi_k) = \eta_k, \quad k=1(1)n,$$

(vgl. Abb. 1a).

Derartige Interpolationsprobleme treten etwa auf, wenn man zu gemessenen Werten  $(\xi_k, \eta_k)$  eine Interpolierende  $\varphi \in \mathfrak{J}$  sucht.

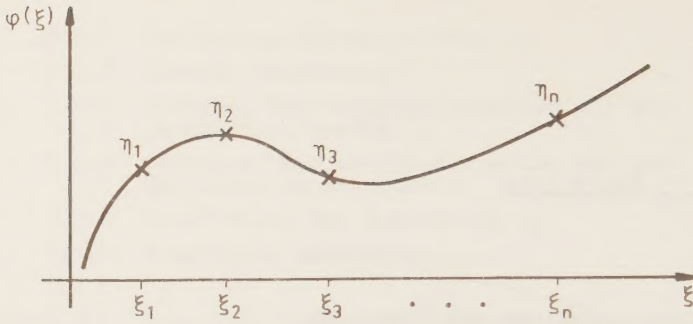


Abb. 1 a Reelles Interpolationsproblem

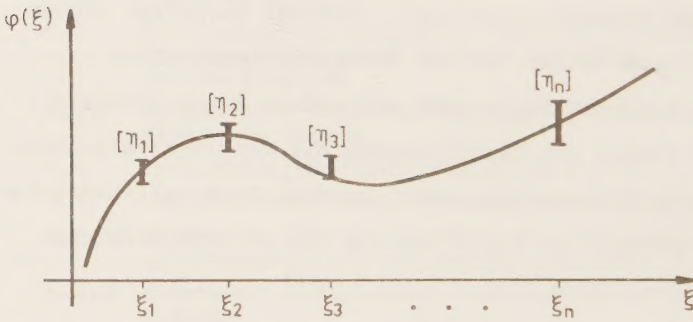


Abb. 1 b Intervallinterpolationsproblem vom Typ I

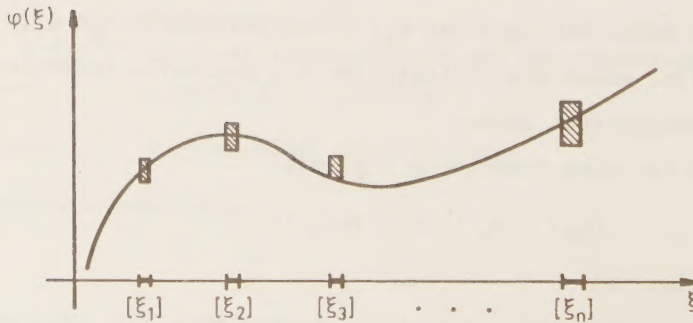


Abb. 1 c Intervallinterpolationsproblem vom Typ II  
mit Rechtecken  $[\xi_k] \times [\eta_k]$

### Intervallinterpolationsproblem

Bei der praktischen Durchführung von Messungen zu verschiedenen Zeitpunkten  $\xi_k$  sind die Meßwerte  $\eta_k$  i.a. infolge von Meßungenauigkeiten zu Werten  $\tilde{\eta}_k \approx \eta_k$  verfälscht. Ursache hierfür sind beispielsweise Ungenauigkeiten der Meßinstrumente und die Ablesungenauigkeit, die aufgrund der Zuordnung eines Resultats etwa eines physikalischen oder chemischen Vorgangs zu einem der beiden nächstgelegenen Skalenwerte entsteht. Oft lassen sich jedoch Schranken  $\underline{\eta}_k, \overline{\eta}_k$  für die Werte  $\eta_k$  angeben, derart daß  $\eta_k \in [\eta_k]$  mit  $[\eta_k] := [\underline{\eta}_k, \overline{\eta}_k]$  gilt. Um beispielsweise Schranken für das Flüssigkeitsniveau in einer Pipette (vgl. Abb. 2) zu bestimmen, nimmt man als untere Schranke  $\underline{\eta}_k$  denjenigen Skalenwert, der nächst unterhalb des Meniskus liegt, und als obere Schranke  $\overline{\eta}_k$  den nächst oberen. Derartige Schranken erhält man auch, wenn die Genauigkeit  $\epsilon$  des Meßvorganges bekannt ist: Setzt man  $[\eta_k] := [\tilde{\eta}_k - \epsilon, \tilde{\eta}_k + \epsilon]$ , dann gilt offensichtlich auch  $\eta_k \in [\eta_k]$ . Derartige Intervalle  $[\eta_k]$  nennen wir auch Meßintervalle.

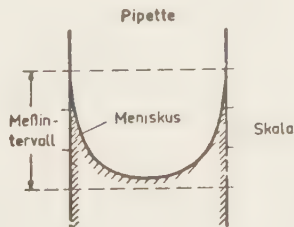


Abb. 2 Flüssigkeitsniveau in einer Pipette mit Meßintervall

Ist nun im Fall der reellen Interpolation eine Funktion  $\varphi \in \mathfrak{F}$  mit  $\varphi(\xi_k) = r_k \in [r_k]$  gesucht, so ist es sinnvoll, sich auf die Bestimmung von Schranken für die Interpolierende  $\varphi$  zu beschränken. Dies führt uns auf das folgende Problem (vgl. Abb. 1b):

Bestimme  $[\varphi] : D \rightarrow II(\mathbb{R})^{1)}$ , derart daß für jedes  $\varphi \in \mathfrak{F}$  mit

$$\varphi(\xi_k) \in [r_k], \quad k=1(1)n,$$

gilt

$$\varphi(\xi) \in [\varphi](\xi) \quad \text{für alle } \xi \in D.$$

Ein Problem dieser Art nennen wir ein Intervallinterpolationsproblem vom Typ I.

Darüberhinaus kann noch angenommen werden, daß von den Stützstellen  $\xi_k$  nur bekannt sei, daß sie in paarweise disjunkten Intervallen  $[\xi_k]$  liegen. Dies kann beispielsweise die Folge ungenauer Zeitmessung sein. Wir erhalten dann ein Intervallinterpolationsproblem vom Typ II (vgl. Abb. 1c):

Bestimme  $[\varphi] : D \rightarrow II(\mathbb{R})$ , derart daß für jedes  $\varphi \in \mathfrak{F}$  mit

$$\varphi(\xi_k) \in [r_k] \quad \text{für (mindestens) ein } \xi_k \in [\xi_k], \quad k=1(1)n,$$

gilt

$$\varphi(\xi) \in [\varphi](\xi) \quad \text{für alle } \xi \in D.$$

Dieses Problem ist allgemeiner als das Problem vom Typ I; mit jeder Lösung des Problems vom Typ II erhalten wir auch eine Lösung des Problems vom Typ I.

Neben der geschilderten Berücksichtigung von Meßungenauigkeiten führen weitere praktische Probleme ebenfalls auf die Intervallinterpolation. So läßt sich beispielsweise

1) Mit  $II(\mathbb{R})$  sei die Menge der kompakten, nichtleeren reellen Intervalle bezeichnet.

die Interpolationsforderung  $\varphi(\xi_k) = \eta_k$ ,  $k=1(1)n$ , des reellen Problems bei der numerischen Berechnung der Interpolierenden  $\varphi$  auf einem Computer wegen des Rundungsfehlereinflusses i.a. nicht erfüllen. Wird dagegen die Berechnung von  $\varphi$  mit einer Maschinenintervallarithmetik (s. Unterabschnitt 3.4.1) durchgeführt, so kann eine Funktion  $[\varphi]$  gefunden werden, die immerhin die Interpolierende  $\varphi$  garantiert einschließt.

#### 1.1.2 Behandlung der Intervallinterpolation in der Literatur

Intervallinterpolationsprobleme wurden schon vereinzelt in der Literatur behandelt: das Problem vom Typ II für algebraische Polynome von Rokne [45] und Majumder/Bhattacharjee [33], Hermite-Interpolation vom Typ II für algebraische Polynome von Bhattacharjee/Majumder [8] und Dobronets/Shokin/Yuldahev [17]. Krawczyk beschäftigte sich in [30] u.a. mit der Fehlerabhängigkeit bei der Interpolation mit algebraischen und trigonometrischen Polynomen (Problem vom Typ I). Das mehrdimensionale Interpolationsproblem vom Typ II untersuchten Bhattacharjee/Majumder in [9].

### 1.1.3 Übersicht

Wir beschäftigen uns in der vorliegenden Arbeit mit Problemen vom Typ I oder II für die Systeme der trigonometrischen Polynome, der Splines und der Funktionen  $\alpha \xi^B$  und  $\alpha_1 \exp(B_1 \xi) + \alpha_2 \exp(B_2 \xi)$ . Dabei werden uns, insbesondere in den Kapiteln 3 und 4, die Interpolationsprobleme auf darüber hinausgehende Probleme führen.

Die vorliegende Arbeit besteht aus fünf Kapiteln, die sich ihrerseits in Abschnitte und Unterabschnitte gliedern.

Im nachfolgenden Abschnitt von Kap.1 notieren wir die für diese Arbeit grundlegenden Bezeichnungen und Definitionen und stellen einige häufig benutzte Hilfsmittel für unsere weiteren Untersuchungen zusammen. Da in späteren Kapiteln zu recht unterschiedlichen Themen auch recht unterschiedliche Hilfsmittel zur Anwendung gelangen, führen wir diese auch erst dort ein, wo sie auch benötigt werden.

In Kap.2 betrachten wir für Probleme vom Typ I den Fall, daß die reelle Interpolationsaufgabe für jede beliebige Vorgabe der Stützwerte eindeutig lösbar ist. In diesem Fall lassen sich sogar optimale Schranken für die Interpolierende angeben.

In Kap.3 führt die Beschäftigung mit der trigonometrischen Interpolation (Problem vom Typ I) auf die Frage nach einer geeigneten intervallarithmetischen Durchführung der

schnellen Fouriertransformation. Die schnelle Fouriertransformation stellt seit ihrer Publizierung im Jahre 1965 durch Cooley/Tuckey [15] ein wichtiges Hilfsmittel für viele Bereiche der Anwendung dar <sup>1)</sup>. Die Ergebnisse von Kap.3 lassen sich daher auch auf Intervall-Probleme, die über die Interpolation hinausgehen, anwenden. Eine derartige Anwendung ist die Einschränkung der Lösungsmenge eines linearen Intervallgleichungssystems; damit beschäftigen wir uns in Zusammenhang mit der Bestimmung von Schranken für die B-Spline-Koeffizienten einer interpolierenden periodischen Spline-Funktion mit äquidistanten Knoten in Unterabschnitt 4.3.1 .

Im zweiten Abschnitt von Kap.3 führen wir die Kreisscheiben- und Rechteckarithmetik ein und stellen einige Eigenschaften komplexer Intervallmatrizen für die folgenden Abschnitte dieses Kapitels zusammen. Die Untersuchungen der Abschnitte 3 und 4 zeigen, daß die Durchführung der schnellen Fouriertransformation in der Kreisscheibenarithmetik bedeutend "engere" Schranken als die Durchführung in der Rechteckarithmetik liefert. Ferner fällt die Schranke für die Akkumulation der Rundungsfehler für die Durchführung der schnellen Fouriertransformation im Fall der Kreisscheibenarithmetik ebenfalls günstiger aus als im Fall der Rechteckarithmetik. Ein Vergleich einer Vielzahl gerechneter Beispiele beschließt dieses Kapitel.

In Kap.4 beschäftigen wir uns mit der Spline-Interpolation und hier vor allem mit Problemen vom Typ II. Da die behandelten Probleme auf Intervallgleichungssysteme führen,

---

<sup>1)</sup> Für eine detaillierte Übersicht siehe beispielsweise [13] und [27].

ist in Abschn.4.1 eine Zusammenstellung weitgehend bekannter Ergebnisse über die Durchführbarkeit des Intervall-Gauß-Algorithmus vorangestellt. In Abschn.4.2 gehen wir zunächst kurz auf die kubische Spline-Interpolation ein. Die sich anschließende Bestimmung von Schranken für die B-Spline-Koeffizienten einer interpolierenden Spline-Funktion führt uns auf das Problem, den von den reellen Matrizen her bekannten Begriff der Totalnichtnegativität auf Intervallmatrizen zu übertragen <sup>1)</sup>. Da die (reellen) totalnichtnegativen Matrizen in vielen Bereichen der Mathematik und Mechanik (vgl. [20] und [29]) eine große Rolle spielen, werden die Untersuchungen der totalnichtnegativen Intervallmatrizen in einem eigenen Abschnitt dieses Kapitels, nämlich Abschn.4.4, durchgeführt. Die wie üblich komponentenweise erklärte Halbordnung auf dem reellen Matrizenraum erweist sich hier als nicht tauglich. Erst wenn diese "schachbrett-transformiert" wird, erhält man einfache Kriterien, mit deren Hilfe von der Totalnichtnegativität der beiden Eckmatrizen einer Intervallmatrix auf die Totalnichtnegativität sämtlicher in dieser Intervallmatrix enthaltener reeller Matrizen geschlossen werden kann. Ferner wird bewiesen: Besitzt ein lineares Intervallgleichungssystem eine Koeffizientenmatrix, die nur reguläre, totalnichtnegative reelle Matrizen enthält, dann kann unter gewissen, leicht nachprüfbaren Bedingungen an die rechte Seite die optimale Intervalleinschließung der

---

1)

Eine reelle Matrix heißt totalnichtnegativ, wenn ihre sämtlichen Unterdeterminanten nichtnegativ sind.

Lösungsmenge dieses Intervallgleichungssystems angegeben werden. An einem Gegenbeispiel wird gezeigt, daß der Intervall-Gauß-Algorithmus abbrechen kann und u.a. aus diesem Grunde diese nicht zu liefern braucht.

In Kap.5 werden schließlich zu den Funktionen  $\sigma \xi^\beta$  Parameterschranken angegeben. Wir zeigen, daß sie optimal sind, und zwar nicht nur für das Intervallinterpolationsproblem vom Typ I, sondern auch vom Typ II. Das Problem für diese spezielle Funktionenklasse wird dazu in ein allgemeineres Problem der Durchschnittsbildung monotoner Funktionsstreifen transformiert und der Beweis der Optimalität für dieses allgemeinere Problem geführt. Ein Ausblick auf die Interpolation mit Exponentialsummen beschließt die vorliegende Arbeit.

#### 1.1.4 Bemerkungen

Es sei noch daraufhingewiesen, daß sich viele Interpolationsverfahren, die hier nicht behandelt werden, mehr oder minder unmittelbar in die Intervallmathematik übertragen lassen, z.B. die Schemata von Aitken/Neville und Newton für die Interpolation mit algebraischen Polynomen. Eine derartige Übertragung beschreiben wir in Abschnitt 4.2 anhand der kubischen Spline-Interpolation.

Das reelle Interpolationsproblem läßt sich formal auf Funktionsstreifen, i.e. Abbildungen  $[f] : D \rightarrow II(\mathbb{R})$ , übertragen: Gesucht ist ein Funktionsstreifen  $[f]$  einer vorgegebenen Klasse von Funktionsstreifen, derart daß die

Gleichheit  $[\omega](\xi_k) = [\gamma_k]$ ,  $k=1(\dots)n$ , gilt. Dieses Problem wurde von Ratschek in [43] für Intervallpolynome untersucht und vollständig gelöst. Wir werden hier nicht auf derartige Probleme eingehen.

Um nicht den Rahmen dieser Arbeit zu sprengen, haben wir darauf verzichtet, vom Reellen her bekannte Ergebnisse aus der Theorie der Interpolation und der Theorie der totalnichtnegativen Matrizen zu beweisen, und uns stattdessen mit Hinweisen auf die Literatur, vor allem da, wo es möglich war, auf leicht zugängliche Lehrbücher begnügt.

Ein Teil der in dieser Arbeit enthaltenen Untersuchungen ist vom Verfasser schon an anderer Stelle (vgl. [21] - [24]), jedoch mit teilweise spezielleren Ergebnissen publiziert worden. Darauf wird im Text häufig nicht mehr hingewiesen werden.

## 1.2 Grundlegende Bezeichnungen, Definitionen und Hilfsmittel

In der vorliegenden Arbeit sind die Formeln u.ä. kapitelweise fortlaufend durchnummeriert. Eine Formel wird durch Angabe des Kapitels, in welchem sie steht, und ihre Nummer zitiert; so bedeutet beispielsweise Formel (4.3): Formel (3) aus Kap. 4. Wird eine Formel im gleichen Kapitel zitiert, so unterdrücken wir die Angabe des Kapitels. Beweise zu Lemmata, Sätzen und Korollaren werden mit  $\blacktriangle$  abgeschlossen.

Dieser Abschnitt enthält Bezeichnungen, Definitionen und elementare Hilfsmittel, die wir im folgenden häufig benötigen werden.

Für Einzelheiten, insbesondere bzgl. der algebraischen Struktur der nachstehend aufgeführten Räume sei auf [2] und [37] verwiesen.

Es sei

$$\left\{ \begin{array}{c} \mathbb{N} \\ \mathbb{Z} \\ \mathbb{R} \\ \mathbb{R}_+ \end{array} \right\} \quad \text{die Menge der} \quad \left\{ \begin{array}{c} \text{natürlichen} \\ \text{ganzen} \\ \text{reellen} \\ \text{positiven reellen} \end{array} \right\} \quad \text{Zahlen.}$$

Reelle Zahlen kennzeichnen wir mit kleinen lateinischen oder kleinen griechischen Buchstaben:  $a, b, c, \dots; \alpha, \beta, \gamma, \dots$ .

$II(\mathbb{R})$  sei die Menge der kompakten, nichtleeren reellen Intervalle. Elemente aus  $II(\mathbb{R})$  notieren wir als  $[a], [b], [c], \dots; [\alpha], [\beta], [\gamma], \dots$  (Vertreter oder Repräsentantenschreibweise) oder auch durch Angabe ihrer Eckpunkte, also beispielsweise

$$[a] = [\underline{a}, \bar{a}] := \{a \in \mathbb{R} \mid \underline{a} \leq a \leq \bar{a}\}.$$

Wie üblich bezeichnen wir für  $[a] \in II(\mathbb{R})$  mit

$$\inf [a] = \underline{a}, \quad \sup [a] = \bar{a}$$

die beiden Eckpunkte des Intervalles  $[a]$ .

Der Einfachheit halber identifizieren wir im folgenden die reellen Zahlen mit den Punktintervallen, das sind Intervalle  $[\underline{a}, \bar{a}]$ , die nur aus einem Punkt bestehen, für die also  $\underline{a} = \bar{a}$  gilt. Zwischen den reellen Zahlen und den Punktintervallen wird im folgenden nicht mehr unterschieden werden.

Ein Intervall  $[a] \in II(\varnothing)$  heißt Nullintervall, falls  $0 \in [a]$  ist. Mit  $]\underline{a}, \overline{a}[$ ,  $]\underline{b}, \overline{b}[$ , ... bezeichnen wir offene, nichtleere reelle Intervalle.

In Erweiterung der reellen Vergleichsoperatoren sei auf  $II(\varnothing)$  wie üblich

$$\left. \begin{aligned} [a] &= [b] : \Leftrightarrow \underline{a} = \underline{b} \wedge \overline{a} = \overline{b}, \\ [a] &\leq [b] : \Leftrightarrow \underline{a} \leq \underline{b} \wedge \overline{a} \leq \overline{b}, \\ [a] &< [b] : \Leftrightarrow \overline{a} < \underline{b}, \end{aligned} \right\} \text{ für alle } [a], [b] \in II(\varnothing),$$

eingeführt. Entsprechend werde  $[a]_{(>)}^>$ ,  $[b]$  verstanden.

Unter dem Vorzeichen eines Intervalles  $[a] \in II(\varnothing)$  verstehen wir

$$\text{sign } [a] := \begin{cases} +1 & \text{falls } 0 < [a] \\ 0 & \text{falls } 0 \in [a] \\ -1 & \text{falls } 0 > [a] \end{cases}.$$

Die arithmetischen Operationen sind erklärt durch

$$[a] \omega [b] := \{a \omega b \mid a \in [a], b \in [b]\}, \omega \in \{+, -, *, /\},$$

wobei im Fall  $\omega = /$  die Division durch ein Nullintervall ausgeschlossen ist. Dies ist äquivalent zu (vgl. etwa [37]):

$$\left. \begin{aligned} [a] + [b] &:= [\underline{a} + \underline{b}, \overline{a} + \overline{b}], \\ [a] - [b] &:= [\underline{a} - \overline{b}, \overline{a} - \underline{b}], \\ [a] * [b] &:= \{\min\{\underline{a}\underline{b}, \underline{a}\overline{b}, \overline{a}\underline{b}, \overline{a}\overline{b}\}, \max\{\underline{a}\underline{b}, \underline{a}\overline{b}, \overline{a}\underline{b}, \overline{a}\overline{b}\}\}, \\ [a] / [b] &:= [\underline{a}, \overline{a}] * [\frac{1}{\underline{b}}, \frac{1}{\overline{b}}], \text{ falls } 0 \notin [b]. \end{aligned} \right\} \quad (1)$$

Im Falle der Multiplikation schreiben wir auch  $[a][b]$  anstelle von  $[a]*[b]$ .

Wie üblich ist

$$[a] \subseteq [b], \text{ falls für alle } a \in [a] \text{ gilt } a \in [b],$$

oder äquivalent

falls  $\underline{b} \leq \underline{a}$  und  $\bar{a} \leq \bar{b}$  ist.

Wir schreiben

$[a] \subset [b]$ , falls die "echte" Inklusion,

d.h. falls  $[a] \subseteq [b]$  mit  $[a] \neq [b]$  gilt.

Für die gesamte Intervallrechnung ist die Eigenschaft

der Inklusionsisotonie der intervallarithmetischen

Operationen grundlegend: Sind  $[a_i], [b_i] \in II(\mathcal{D})$  mit

$[a_i] \subseteq [b_i]$ ,  $i=1,2$ , und ist  $\omega \in \{+, -, *, /\}$ , dann gilt

$$[a_1] \omega [a_2] \subseteq [b_1] \omega [b_2] .$$

Wir benötigen die Abbildungen  $\text{mig}, |\cdot|, \text{span}: II(\mathcal{D}) \rightarrow \mathcal{D} \cup \{0\}$ :

$\text{mig}[a] := \min\{|a| \mid a \in [a]\}$  , die Mignitude ,

$|[a]| := \max\{|a| \mid a \in [a]\}$  , der Betrag ,

$\text{span}[a] := \bar{a} - \underline{a}$  , die Spanne ,

des Intervalls  $[a] \in II(\mathcal{D})$  .

Wir notieren einige Rechenregeln für die Spanne und den

Betrag (vgl. [2], Abschn. 2):

Lemma 1.1: Für  $[a], [b] \in II(\mathcal{D})$ ,  $\alpha \in \mathcal{R}$ , gilt

$$[a] \subseteq [b] \Rightarrow |[a]| \leq |[b]|, \text{span}[a] \leq \text{span}[b], \quad (2a)$$

$$|[b]| \leq |[a]| + \text{span}[b] - \text{span}[a] \quad (2b)$$

$$|[a]| \geq 0 \text{ und } |[a]| = 0 \Leftrightarrow [a] = 0$$

$$|[a] + [b]| \leq |[a]| + |[b]| \quad (2c)$$

$$|\alpha[a]| = |\alpha| |[a]|$$

$$\text{span}([a] \pm [b]) = \text{span}[a] + \text{span}[b]$$

$$\text{span}(\alpha[a]) = |\alpha| \text{span}[a]$$

$$\text{span}([a][b]) = |[b]| \text{span}[a], \text{ falls } [a] = -[a] \text{ gilt,} \quad (2d)$$

$$\text{span}([a][b]) \geq \max\{|[a]| \text{span}[b], |[b]| \text{span}[a]\} \quad (2e)$$

Beweis: Die Aussage (2b) konnte in der intervallmathematischen Literatur nicht gefunden werden und soll nachstehend bewiesen werden.

Wegen  $\bar{a} \leq \bar{b}$  ist  $\underline{a} - (\bar{b} - \underline{b} - (\bar{a} - \underline{a})) \leq \underline{b}$ ,  
wegen  $\underline{b} \leq \underline{a}$  gilt  $\bar{b} \leq \bar{a} + (\bar{b} - \underline{b} - (\bar{a} - \underline{a}))$ ,  
und es folgt

$$[b] \subseteq [a] + (\text{span}[b] - \text{span}[a]) [-1, 1]$$

bzw. wegen (2a) und (2c)

$$|[b]| \leq |[a]| + |\text{span}[b] - \text{span}[a]| [-1, 1],$$

woraus (2b) folgt. ▲

Mit  $\mathcal{R}^n$  sei der Raum der  $n$ -dimensionalen Vektoren mit Komponenten aus  $\mathcal{R}$  bezeichnet.

Ist  $\Omega$  eine nichtleere Menge, dann bezeichnen wir mit  $M_{n,m}(\Omega)$  die Menge der  $n \times m$ -Matrizen, deren Komponenten Elemente aus  $\Omega$  sind. Ist aus dem Kontext ersichtlich, welche Menge  $\Omega$  gemeint ist, schreiben wir auch abkürzend  $M_{n,m}$  anstelle von  $M_{n,m}(\Omega)$ . Für  $m=1$  erhalten wir die Menge der  $n$ -Tupel über  $\Omega$ . Viele Operationen, die wir hier und auch später für Matrizen formulieren werden, sind damit auch für  $n$ -Tupel gültig.

Insbesondere bezeichnet  $M_{n,m}(\mathcal{R})$  bzw.  $M_{n,m}(\text{II}(\mathcal{R}))$  die Menge der  $n \times m$ -Matrizen über  $\mathcal{R}$  bzw.  $\text{II}(\mathcal{R})$ . Elemente aus  $M_{n,m}(\mathcal{R})$  bezeichnen wir mit kleinen deutschen Buchstaben  $a = (a_{ij})$ ,  $b = (b_{ij}), \dots$ , Elemente aus  $M_{n,m}(\text{II}(\mathcal{R}))$  mit

$$[a] = ([a_{ij}]), [b] = ([b_{ij}]), \dots \quad (3)$$

Mit  $a^t$  bezeichnen wir die Transponierte zu  $a$ . Unter einer

Diagonalmatrix  $\text{diag}(a_1, \dots, a_n)$  verstehen wir wie  
 üblich eine Matrix, die höchstens auf der Hauptdiago-  
 nalen nichtverschwindende Koeffizienten besitzt und deren  
 Koeffizienten auf der Hauptdiagonalen  $a_1, \dots, a_n$  sind.  
 Der Einfachheit halber identifizieren wir  $M_{n,n}(\mathfrak{P})$  mit  
 dem Raum der  $n \times m$ -Punktmatrizen, i.e. die Elemente  
 $[a] = ([a_{ij}]) \in M_{n,m}(\text{II}(\mathfrak{P}))$ , deren Komponenten sämtlich  
 Punktintervalle sind; zwischen den Elementen aus  $M_{n,m}(\mathfrak{P})$   
 und den  $n \times m$ -Punktmatrizen wird im folgenden nicht mehr  
 unterschieden werden.

Zu  $[a] \in M_{n,m}(\text{II}(\mathfrak{P}))$  sei  $\text{sign}[a] \in M_{n,m}(\mathfrak{P})$  gegeben durch

$$(\text{sign}[a])_{ij} := \text{sign}[a_{ij}], \quad i=1(1)n, \quad j=1(1)m.$$

Auf  $M_{n,m}(\text{II}(\mathfrak{P}))$  werden die Addition, Subtraktion und  
 Multiplikation formal wie bei den Punktmatrizen komponenten-  
 tenweise erklärt

$$([a] \pm [b])_{ij} := [a_{ij}] \pm [b_{ij}], \quad i=1(1)n, \quad j=1(1)m, \\ \text{für } [a], [b] \in M_{n,m},$$

$$([a] * [b])_{ij} := \bigwedge_{u=1}^m [a_{iu}] [b_{uj}], \quad i=1(1)n, \quad j=1(1)m, \\ \text{für } [a] \in M_{n,m}, [b] \in M_{m,r}.$$

Für  $[a] * [b]$  schreiben wir wieder  $[a][b]$ . Es gilt

$$[a] \pm [b] = \{a \pm b \mid a \in [a], b \in [b]\}. \quad (4)$$

Auf  $M_{n,m}(\mathfrak{P})$  sei die Halbordnung  $\leq$  komponentenweise er-  
 klärt gemäß

$$a \leq b : \Leftrightarrow a_{ij} \leq b_{ij}, \quad i=1(1)n, \quad j=1(1)m, \\ \text{für } a, b \in M_{n,m}.$$

Mit dieser Halbordnung wird  $M_{n,m}(\mathcal{O})$  zu einem bedingt vollständigen Vektorverband (zur Begriffsbildung s. etwa [48]).

Ferner gilt für  $[a] = ([a_{ij}, \bar{a}_{ij}]) \in M_{n,m}(II(\mathcal{O}))$

$$[a] = [\underline{a}, \bar{a}] \text{ mit } \underline{a} = (\underline{a}_{ij}), \bar{a} = (\bar{a}_{ij}). \quad (5)$$

In (5) ist  $[a]$  dargestellt als Element von  $II(M_{n,m}(\mathcal{O}))$ , des Raumes der Intervalle (bzgl. der Halbordnung  $\leq$ ) über  $M_{n,m}(\mathcal{R})$ . Wir werden im folgenden wie üblich die Redeweisen "Intervallmatrix" (als Element von  $M_{n,m}(II(\mathcal{O}))$ ) und "Matrix-Intervall" (als Element von  $II(M_{n,m}(\mathcal{O}))$ ) sowie die Bezeichnungen (3) und (5) nebeneinander verwenden.

Im Fall  $m=1$  schreiben wir für  $M_{n,1}(II(\mathcal{O}))$  bzw.  $II(M_{n,1}(\mathcal{O}))$  auch  $II(\mathcal{O}^n)$  und sprechen auch von Intervallvektoren.

Auf  $M_{n,m}(II(\mathcal{Q}))$  wird  $(\leq)$  komponentenweise erklärt gemäß

$$[a] \leq [b] : \Leftrightarrow [a_{ij}] \leq [b_{ij}], i=1(\dots)n, j=1(1)m. \quad (6)$$

Für eine beschränkte, nichtleere Teilmenge  $\mathcal{Q}$  des  $M_{n,m}(\mathcal{O})$  heißt

$$\text{Intv } \mathcal{Q} := [\inf \mathcal{Q}, \sup \mathcal{Q}]$$

die optimale Intervalleinschließung [4] oder Hülle [5] von  $\mathcal{Q}$ .

Die optimale Intervalleinschließung läßt sich komponentenweise bestimmen:

Für  $i=1(1)n, j=1(1)m$  bezeichne  $\pi_{ij}$  die  $(i,j)$ -te Projektion des  $M_{n,m}(\mathcal{R})$  nach  $\mathcal{R}$ ,

$$\pi_{ij}(a) := a_{ij} \text{ für } a \in M_{n,m}(\mathcal{O}).$$

Lemma 1.2: Ist  $\mathcal{Q}$  eine nichtleere, beschränkte Teilmenge des  $M_{n,m}(\mathbb{R})$ , dann gilt

$$\pi_{ij}(\text{Intv } \mathcal{Q}) = \text{Intv } \pi_{ij}(\mathcal{Q}), \quad i=1(\wedge)n, \quad j=1(\wedge)m.$$

Beweis: Da die Halbordnung komponentenweise erklärt ist, gilt

$$\begin{aligned} \text{Intv } \mathcal{Q} &= [\inf \mathcal{Q}, \sup \mathcal{Q}] = ([\inf \pi_{ij}(\mathcal{Q}), \sup \pi_{ij}(\mathcal{Q})])_{\substack{i=1(\wedge)n, \\ j=1(\wedge)m}} \\ &= (\text{Intv } \pi_{ij}(\mathcal{Q}))_{\substack{i=1(\wedge)n, \\ j=1(\wedge)m}}. \quad \blacktriangle \end{aligned}$$

Als Folgerung erhalten wir in Ergänzung zu (4) Satz 60 aus [40], S. 159.

Korollar 1.1: Für  $[a] \in M_{n,m}(\text{II}(\mathcal{D}))$ ,  $[b] \in M_{m,r}(\text{II}(\mathcal{D}))$  ist

$$\text{Intv}\{a \cdot b \mid a \in [a], b \in [b]\} = [a][b].$$

Lemma 1.3: Es seien  $\mathcal{Q}_i$ ,  $i=1(\wedge)k$ , nichtleere, beschränkte Teilmengen von  $M_{n,m}(\mathcal{D})$  und es sei  $\bigcap_{i=1}^k \mathcal{Q}_i \neq \emptyset$ . Dann gilt

$$\text{Intv}(\bigcap_{i=1}^k \mathcal{Q}_i) \subseteq \bigcap_{i=1}^k \text{Intv} \mathcal{Q}_i.$$

Beweis: Es ist  $\mathcal{Q}_i \subseteq \text{Intv} \mathcal{Q}_i$  und somit

$$\bigcap \mathcal{Q}_i \subseteq \bigcap \text{Intv} \mathcal{Q}_i$$

$$\text{Intv } \bigcap \mathcal{Q}_i \subseteq \text{Intv}(\bigcap \text{Intv} \mathcal{Q}_i) = \bigcap \text{Intv} \mathcal{Q}_i. \quad \blacktriangle$$

Bemerkung: I.a. gilt nicht die Gleichheit, wie das folgende Beispiel zeigt: Für  $\epsilon \in \mathbb{R}_+$  und  $(x_1, x_2) \in \mathbb{R}^2$  sei definiert

$$K_{\epsilon}^1(x_1, x_2) := \{(y_1, y_2) \in \mathbb{R}^2 \mid |y_1 - x_1| + |y_2 - x_2| \leq \epsilon\},$$

$$K_{\epsilon}^{\infty}(x_1, x_2) := \{(y_1, y_2) \in \mathbb{R}^2 \mid \max\{|y_1 - x_1|, |y_2 - x_2|\} \leq \epsilon\}.$$

Es ist

$$\text{Intv } K_{\epsilon}^1(x_1, x_2) = K_{\epsilon}^{\infty}(x_1, x_2),$$

$$K_1^1(0,0) \cap K_1^1(1,0) = K_{0.5}^1(0.5,0),$$

$$\text{Intv } K_1^1(0,0) \cap \text{Intv } K_1^1(1,0) = [0,1] \times [-1,1]$$

$$\supset [0,1] \times [-0.5,0.5] = \text{Intv } K_{0.5}^1(0.5,0).$$

Wie schon in Unterabschn. 1.1.4 angemerkt, verstehen wir unter einem Funktionsstreifen eine Abbildung

$$[\varphi] : D \rightarrow \Pi(\mathbb{R}), \quad D \subseteq \mathbb{R}. \quad (7)$$

Derartige Abbildungen notieren wir auch durch Angabe ihrer Randfunktionen

$$[\varphi] = [\underline{\varphi}, \overline{\varphi}] \quad \text{mit}$$

$$\underline{\varphi}(x) := \inf[\varphi](x), \quad \overline{\varphi}(x) := \sup[\varphi](x).$$

Es sei  $\varphi : D \rightarrow \mathbb{R}$ . Dann heit  $[\varphi]$  aus (7) eine Funktionsstreifeneinschlieung zu  $\varphi$ , falls gilt

$$\varphi(x) \in [\varphi](x) = [\underline{\varphi}(x), \overline{\varphi}(x)] \quad \text{fr alle } x \in D.$$

Eine reelle Funktion  $\varphi : D \rightarrow \mathbb{R}$ ,  $D \subseteq \mathbb{R}$ , heit

$$\left\{ \begin{array}{l} \text{isoton} \\ \text{antiton} \end{array} \right\}, \quad \text{wenn aus } x_1 \leq x_2 \text{ folgt } \left\{ \begin{array}{l} \varphi(x_1) \leq \varphi(x_2) \\ \varphi(x_1) \geq \varphi(x_2) \end{array} \right\}$$

fr alle  $x_1, x_2 \in D$ ,

$$\text{streng } \left\{ \begin{array}{l} \text{isoton} \\ \text{antiton} \end{array} \right\}, \quad \text{wenn aus } x_1 < x_2 \text{ folgt } \left\{ \begin{array}{l} \varphi(x_1) < \varphi(x_2) \\ \varphi(x_1) > \varphi(x_2) \end{array} \right\}$$

fr alle  $x_1, x_2 \in D$ ,

<streng> monoton , wenn sie isoton oder antiton  
<im strengen Sinne> ist.

Wir nennen einen Funktionsstreifen <streng> isoton  
bzw. <streng> antiton , falls seine beiden Randfunk-  
tionen die entsprechende Eigenschaft besitzen.

## Kapitel 2 Optimalität der Einschließung

Wir beschränken uns in diesem Kapitel auf Interpolationsprobleme vom Typ I. Für Probleme vom Typ II kann mit den nachstehend aufgeführten Methoden i.a. keine Optimalität erzielt werden.

Ein Problem vom Typ I bedeutet, daß Schranken für eine unbekannte Interpolierende  $\varphi \in \mathfrak{F}$  mit  $\varphi(\xi_k) = \eta_k$  und  $\eta_k \in [\eta_k]$  zu ermitteln sind. Damit sind aber auch Schranken für jede Interpolierende  $\varphi$  mit Wertevorgabe  $\eta_k$ ,  $\eta_k \in [\eta_k]$ , und damit für eine ganze Lösungsfamilie zu bestimmen.

### 2.1 Optimale Funktionsstreifeneinschließung

Definition (<optimale> Funktionsstreifeneinschließung der Lösungsfamilie):

Wir nennen einen Funktionsstreifen  $[\varphi] : D \rightarrow II(\mathbb{R})$  eine Funktionsstreifeneinschließung der Lösungsfamilie, wenn  $[\varphi]$  Funktionsstreifeneinschließung jedes Elementes der Lösungsfamilie ist, d.h. falls

$$\left. \begin{array}{l} \text{aus } \varphi \in \mathfrak{F} \text{ und } \varphi(\xi_k) \in [\eta_k], \quad k=1(1)n, \\ \text{folgt } \varphi(\xi) \in [\varphi](\xi) \text{ für alle } \xi \in D. \end{array} \right\} \quad (1)$$

Eine Funktionsstreifeneinschließung  $[\varphi]$  der Lösungsfamilie heißt optimal, wenn sie in jedem Funktionsstreifen  $[\varphi]$  mit der Einschließungseigenschaft (1) enthalten ist.

Wir betrachten nun den Sonderfall, daß das reelle Interpolationsproblem für jede Wertevorgabe eindeutig lösbar ist.

Es sei  $\mathfrak{F}$  ein reeller  $n$ -dimensionaler Vektorraum von Funktionen  $\varphi : D \rightarrow \mathbb{R}$  mit dem unendlich viele Elemente enthaltenden, zusammenhängenden Definitionsbereich  $D \subseteq \mathbb{R}$ .

Es gelte ferner:

$$\left. \begin{array}{l} \text{Zu beliebig vorgegebenen paarweise verschiedenen} \\ \xi_k \in D, k=1(1)n, \text{ und beliebigen } \eta_k \in \mathbb{R}, k=1(1)n, \\ \text{gibt es eine Funktion } \varphi \in \mathfrak{F} \text{ mit} \\ \varphi(\xi_k) = \eta_k, \quad k=1(1)n. \end{array} \right\} (2)$$

Aus der Lösbarkeit der reellen Interpolationsaufgabe ergibt sich schon die Eindeutigkeit der Interpolierenden  $\varphi$  (vgl. [53], S.46).

In [53], Kap. I, § 4, sind einige Beispiele für derartige Funktionensysteme aufgeführt. Es sei hier vor allem auf die algebraischen und trigonometrischen Polynome und Funktionen (vgl. [20], S.88)

$$\left. \begin{array}{l} \gamma_1 \xi^{\alpha_1} + \gamma_2 \xi^{\alpha_2} + \dots + \gamma_n \xi^{\alpha_n} \\ \text{mit gegebenen } \alpha_1 < \alpha_2 < \dots < \alpha_n \text{ und } D := \mathbb{R}_+ \end{array} \right\} (3)$$

hingewiesen.

Zu den  $n^2$  Wertevorgaben  $\delta_{kj}$  (Kroneckersymbol),  $k, j=1(1)n$ , existieren nach (2) gerade  $n$  Funktionen  $\varphi_k \in \mathfrak{F}$ ,  $k=1(1)n$ , mit

$$\varphi_k(\xi_j) = \delta_{kj}, \quad j=1(1)n.$$

Damit genügt für eine beliebige Wertevorgabe

$\eta_1, \eta_2, \dots, \eta_n \in \mathbb{R}$  die Funktion  $\varphi$ ,

$$\varphi(\xi) := \sum_{k=1}^n \eta_k \varphi_k(\xi) \quad , \quad \xi \in D \quad , \quad (5)$$

der Interpolationsforderung

$$\varphi(\xi_j) = \eta_j \quad , \quad j=1(1)n \quad .$$

Satz 2.1: Im Fall der eindeutigen Lösbarkeit des reellen Interpolationsproblems ist der Funktionsstreifen  $[\psi]$ ,

$$[\psi](\xi) := \sum_{k=1}^n [\eta_k] \varphi_k(\xi) \quad , \quad \xi \in D \quad , \quad (6)$$

die optimale Funktionsstreifeneinschließung der Lösungsfamilie des Intervallinterpolationsproblems vom Typ I.

Beweis: Die Einschließungsaussage folgt aus der Inklusionsisotonie der intervallararithmetischen Operationen. In (5) tritt jede der Größen  $\eta_k$  genau einmal auf. Damit ergibt sich die Optimalitätsaussage des Satzes unmittelbar aus der Optimalitätsaussage in [37], S.11.  $\blacktriangle$

Bemerkung: Der Spezialfall der Interpolation mit algebraischen Polynomen wurde in [28] und [22] behandelt.

Die von der Stützstelle  $\xi_k$  verschiedenen Stützstellen  $\xi_j$  sind die einzigen Nullstellen der Funktion  $\varphi_k$  auf  $D$  (vgl. [53], S.41). Sind die Funktionen aus  $\mathfrak{F}$  sämtlich stetig, dann ist damit das Vorzeichen jeder Funktion  $\varphi_k$  auf  $[\xi_{j-1}, \xi_j]$  konstant. Für viele bekannte Funktionensysteme z.B. für algebraische und trigonometrische

Polynome lassen sich die Vorzeichenverteilungen der Funktionen  $\varphi_k$  und damit auch die Randfunktionen  $\underline{\psi}, \overline{\psi}$  von  $[\psi]$  explizit angeben. Insbesondere setzen sich die beiden Randfunktionen stückweise aus Funktionen aus  $\mathfrak{F}$  mit der Interpolationseigenschaft bzgl. bestimmter Randpunkte der Intervalle  $[\eta_k]$  zusammen.

Es werde nun angenommen, daß gilt

$$\text{span}[\eta_k] \leq \alpha, \quad k=1(1)n.$$

Damit ist nach geläufigen Regeln für das Rechnen mit der Spanne (vgl. Lemma 1.1) für alle  $\xi \in D$

$$\text{span}[\psi](\xi) = \sum_{k=1}^n \text{span}[\eta_k] |\varphi_k(\xi)| \leq \alpha s_{[\psi]}(\xi)$$

mit

$$s_{[\psi]}(\xi) := \sum_{k=1}^n |\varphi_k(\xi)|.$$

Die Funktion  $s_{[\psi]}$  ist die von reellen Interaktionsprozessen her bekannte Lebesgue-Funktion bzgl.

$\varphi_1, \varphi_2, \dots, \varphi_n$ . Eine Übersicht über Abschätzungen der sog.

$$\text{Lebesgue-Konstante} \quad \max_{\xi \in D} s_{[\psi]}(\xi)$$

für polynomiale Funktionensysteme befindet sich beispielsweise in [14].

## 2.2 Optimale Intervalleinschließung der Parameter der Lösungsfamilie

In der Praxis wird die gegebene Funktionenklasse  $\mathfrak{F}$  oft

durch eine parameterabhängige Funktion

$$\left. \begin{aligned} \varphi(\xi; \gamma_1, \dots, \gamma_m) : D \rightarrow \mathbb{R}, \\ D \subseteq \mathbb{R}, \gamma_i \in \mathbb{R}, i=1(1)m, \end{aligned} \right\} \quad (7)$$

erzeugt. Beispiele hierfür sind wieder die Polynome und Funktionen aus (3). Häufig läßt sich das vorgelegte Problem vom Typ 1, d.h. Bestimmung eines Funktionsstreifens  $[\varphi]$ , der die Interpolierende  $\varphi$  enthält, lösen, indem man Schranken für die Parameter  $\gamma_i$  der Interpolierenden  $\varphi$  bestimmt. Hat man nämlich Intervalle  $[\gamma_i]$  mit  $\gamma_i \in [\gamma_i]$  ermittelt und führt die Ersetzung der  $\gamma_i$  durch die Intervalle  $[\gamma_i]$  in (7) zu einem definierten Intervallausdruck (vgl. [2], S.28-30), dann liefert wegen der Inklusionsisotonie der intervallarithmetrischen Operationen die Intervallerweiterung  $[\varphi](\xi; [\gamma_1], \dots, [\gamma_m])$  eine Funktionsstreifeneinschließung von  $\varphi$ . Beispiele hierzu enthalten die Kapitel 3,4 und 5.

Definition (optimale Intervalleinschließung der Parameter der Lösungsfamilie):

Wir nennen den Intervallvektor  $[\alpha] = ([\gamma_i])_{i=1}^m$  optimale Intervalleinschließung der Parameter der Lösungsfamilie, falls gilt

$$[\alpha] = \text{Intv}\{(\gamma_1, \dots, \gamma_m)^t \in \mathbb{R}^m \mid \varphi(\xi_k; \gamma_1, \dots, \gamma_m) \in [\eta_k], k=1(1)m\} \quad (8)$$

Wir betrachten nun den Spezialfall von (7), in dem  $m=n$  gilt,  $\mathfrak{X}$  wieder ein  $n$ -dimensionaler Vektorraum ist und die Parameter  $\gamma_1, \dots, \gamma_n$  von  $\varphi$  gerade die Koeffizienten von  $\varphi$  bzgl. der Basis  $\{\varphi_1, \dots, \varphi_n\}$  von  $\mathfrak{X}$  sind, d.h.

$$\varphi(\xi; v_1, \dots, v_n) = \sum_{i=1}^n v_i \varphi_i(\xi) \quad , \quad \xi \in D \quad .$$

Die Interpolationsaufgabe sei wieder (eindeutig) lösbar, d.h. es gelte (2). Dann führt die Interpolationsforderung  $\varphi(\xi_k; v_1, \dots, v_n) \in [\eta_k]$  auf das lineare Gleichungssystem

$$\begin{pmatrix} \varphi_1(\xi_k) \\ \vdots \\ \varphi_n(\xi_k) \end{pmatrix} * \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix} \quad \text{mit } \eta_k \in [\eta_k], \quad k=1(1)n \quad .$$

Die Koeffizientenmatrix dieses linearen Gleichungssystems ist wegen der Lösbarkeit der reellen Interpolationsaufgabe regulär (vgl. [53], S.41). Nach Korollar 1.1 ist dann die optimale Intervalleinschließung der Parameter der Lösungsfamilie gegeben durch den Intervallvektor  $[a]$

$$\left. \begin{aligned} [a] &:= (\varphi_i(\xi_k))^{-1} [\eta] \\ \text{mit } [\eta] &= ([\eta_k])_{k=1}^n \end{aligned} \right\} \quad (9)$$

Wir weisen noch daraufhin, daß die oben geschilderte Ersetzung der Parameter  $v_i$  durch die Komponenten  $[v_i]$  der optimalen Intervalleinschließung der Parameter der Lösungsfamilie i.a. nicht zu einer optimalen Funktionsstreifeneinschließung führt.

In Kap. 5 werden wir mit anderen Mitteln zeigen, daß die dort für die Klasse der Funktionen  $\sigma \xi^B$  hergeleiteten Schranken für die Parameter  $\sigma$  und  $B$  optimal sind, und zwar sogar für die Intervallinterpolation vom Typ II.

Die Einschließung, die wir in diesem Kapitel angegeben haben, sind vor allem vom Standpunkt der Theorie her interessant.

- 4 -

Für die Funktionensysteme der trigonometrischen Polynome und der kubischen Splines, insbesondere wenn die Schranken für mehrere Argumente berechnet werden sollen, sind sie jedoch mit größerem Aufwand zu berechnen als die (i.a. nicht optimalen) Einschließungen, die wir in den Kapiteln 3 und 4 angeben werden.

Für die Hermite-Interpolation läßt sich mit ähnlichen Mitteln Optimalität sowohl der Funktionsstreifeneinschließung als auch der Einschließung der Parameter der Lösungsfamilie erzielen.

## Kapitel 3 Intervallinterpolation mit trigonometrischen Polynomen

### 3.1 Vorbemerkungen

In diesem Kapitel werden wir uns der komplexen Rechnung bedienen. Um die komplexen von den reellen Zahlen zu unterscheiden, bezeichnen wir in diesem Kapitel die reellen Zahlen mit kleinen griechischen Buchstaben (Ausnahme: Indizes) und die komplexen Zahlen mit kleinen lateinischen Buchstaben:  $a = \alpha^{(1)} + i \alpha^{(2)}$ ,  $x = \xi + i\eta, \dots$  ( $i := \sqrt{-1}$ ). Mit  $\bar{a}$  werde das konjugiert-komplexe zu  $a$  bezeichnet.

Bekanntlich (vgl. etwa [50], S. 6') gibt es zu den  $2n$  Stützstellen  $\xi_k$ ,  $\xi_k := k\pi/n$ , und Stützwerten  $\eta_k \in \mathbb{R}$ ,  $k=0(1)2n-1$ ,  $n \geq 1$ , genau ein trigonometrisches Polynom  $\varphi(\xi)$ ,

$$\begin{aligned} \varphi(\xi) &= \varphi(\xi; \alpha_0, \dots, \alpha_n, \beta_1, \dots, \beta_{n-1}) := \\ &= \frac{\alpha_0}{2} + \sum_{j=1}^{n-1} \alpha_j \cos j\xi + \beta_j \sin j\xi + \frac{\alpha_n}{2} \cos n\xi, \end{aligned} \quad (1)$$

das der Interpolationsforderung

$$\varphi(\xi_k) = \eta_k, \quad k=0(1)2n-1,$$

genügt.<sup>1)</sup> Die Koeffizienten  $\alpha_j, \beta_j$  sind gegeben durch

---

) In Hinblick auf unsere nachfolgenden Untersuchungen zur intervallarithmetischen Durchführung der sog. schnellen Fouriertransformation beschränken wir uns hier bewußt auf diese speziellen trigonometrischen Polynome.

$$\left. \begin{aligned} \alpha_j &:= \frac{1}{n} \sum_{k=0}^{2n-1} \eta_k \cos j \varepsilon_k, \quad j = 0(1)n, \\ \beta_j &:= \frac{1}{n} \sum_{k=0}^{2n-1} \eta_k \sin j \varepsilon_k, \quad j = 1(1)n-1. \end{aligned} \right\} \quad (2)$$

In diesem Kapitel soll das Intervallinterpolationsproblem vom Typ I behandelt werden. Wegen der (eindeutigen) Lösbarkeit des reellen Interpolationsproblems kann mit der Vorgehensweise aus Abschn. 2.1 die optimale Funktionsstreifeneinschließung der Lösungsfamilie bestimmt werden. Um die Spanne dieser Einschließung gemäß der Vorgehensweise in Abschnitt 2.1 nach oben hin abschätzen zu können, geben wir in Unterabschnitt 3.1.1 die Lebesgue-Konstante für die trigonometrischen Polynome an. Von Wichtigkeit für die Anwendungen ist die Bestimmung von Schranken für die Koeffizienten  $\alpha_j$  und  $\beta_j$ , d.h. für Intervalle  $[\alpha_j], [\beta_j]$ , für die gilt

$$\left. \begin{aligned} \varphi(\varepsilon_k; \alpha_0, \dots, \alpha_n, \beta_1, \dots, \beta_{n-1}) &\in [\eta_k], \quad k=0(1)2n-1, \\ \rightarrow \quad \left\{ \begin{aligned} \alpha_j &\in [\alpha_j], \quad j=0(1)n, \\ \beta_j &\in [\beta_j], \quad j=1(1)n-1. \end{aligned} \right. \end{aligned} \right\} \quad (3)$$

Ersetzt man dazu in (2) die reellen Zahlen  $\eta_k$  durch die Intervalle  $[\eta_k]$ , so liefern nach Korollar 1.1, die Intervalle  $[\alpha_j^*], [\beta_j^*]$ ,

$$\left. \begin{aligned} [\alpha_j^*] &:= \frac{1}{n} \sum_{k=0}^{2n-1} [\eta_k] \cos j \varepsilon_k, \quad j=0(1)n, \\ [\beta_j^*] &:= \frac{1}{n} \sum_{k=0}^{2n-1} [\eta_k] \sin j \varepsilon_k, \quad j=1(1)n-1, \end{aligned} \right\} \quad (2')$$

die optimalen Intervalleinschlüsse der Koeffizienten  $\alpha_j, \beta_j$  der trigonometrischen Interpolationspolynome. Der Arbeitsaufwand zur Bestimmung dieser optimalen Intervalleinschlüsse beträgt  $O(n^2)$  Operationen. Wir wollen nun annehmen, daß  $n$  eine Potenz von 2 ist, etwa  $n = 2^p$  mit  $p$  (positive) natürliche Zahl. Für diese Wahl von  $n$  kann die sog. schnelle Fouriertransformation, im folgenden mit FFT (  $\hat{=}$  Fast Fourier Transform ) abgekürzt, angewendet werden.<sup>1)</sup> Diese Transformation kann, wie wir zeigen werden, auch intervallmäßig durchgeführt werden. Man erhält dann jedoch nicht mehr die optimalen Intervalleinschlüsse  $[\alpha'_j], [\beta'_j]$  aus (2'), sondern nur noch Intervalle  $[\alpha_j], [\beta_j]$ , die diese Koeffizienten enthalten. In Abschn. 3.3 werden wir uns u.a. mit der Frage befassen, wie groß die Überschätzung von  $[\alpha'_j]$  bzw.  $[\beta'_j]$  durch die Intervalle  $[\alpha_j]$  bzw.  $[\beta_j]$  ausfallen kann. Bei Einsatz der FFT verringert sich jedoch der Arbeitsaufwand auf  $O(np)$  Operationen. Das bedeutet, daß für große  $n$  der Berechnung der optimalen Intervalleinschlüsse  $[\alpha'_j], [\beta'_j]$  gegenüber nicht nur die benötigte Rechenzeit wesentlich vermindert wird, sondern auch, daß eine Aufblähung der mit einer Maschinenintervallarithmetik berechneten Einschließungsintervalle  $[\alpha'_j], [\beta'_j]$  weitgehend vermieden wird (vgl. Unterabschn. 3.4.6).

---

1) Eine detaillierte Darstellung der schnellen Fouriertransformation nebst einer umfangreichen Bibliographie enthält [13].

Bevor wir uns der FFT zuwenden, werden wir uns zunächst in Unterabschn. 3.1.2 mit der diskreten Fouriertransformation befassen. Um die diskrete (und damit auch die schnelle) Fouriertransformation intervallmäßig durchführen zu können, benötigen wir noch einige Hilfsmittel, die in Abschn. 3.2 bereitgestellt werden.

### 3.1.1 Die Lebesgue-Konstante

Setzt man (2) in (1) ein und vertauscht die Summationen, dann erhält man die Funktion  $\varphi_k$  mit (2.4):

$$\varphi_k(\xi) = \frac{1}{2n} \left( 1 + 2 \sum_{j=1}^{n-1} \cos j(\xi_k - \xi) + (-1)^k \cos n\xi \right).$$

Wir notieren einige Eigenschaften der Funktionen  $\varphi_k$  (vgl. [18]):

Für  $\xi \neq \xi_k$  läßt  $\varphi_k$  sich darstellen als

$$\varphi_k(\xi) = \frac{1}{2n} (-1)^k \sin n\xi \cot \frac{1}{2} (\xi - \xi_k).$$

Die Nullstellen von  $\varphi_k$  auf  $[0, 2\pi[$  sind genau die von der Stützstelle  $\xi_k$  verschiedenen Stützstellen  $\xi_j$ ;  $\varphi_k'$  verschwindet bei  $\xi_j$  genau dann, wenn  $|k-j| = n$  ist. Damit ist die Vorzeichenverteilung der Funktionen  $\varphi_k$  auf den Intervallen  $[\xi_{j-1}, \xi_j[$  festgelegt, und die Randfunktionen der optimalen Funktionsstreifeneinschließung [4] nach (2.6) lassen sich über jedem dieser Teilintervalle explizit angeben.

Für die Lebesgue-Funktion bzgl.  $\varphi, \dots, \varphi_n$  gilt:  
 $s_{[\psi]}$  hat die Periode  $\pi/n$ , ist symmetrisch bzgl.  $\pi/2n$   
 und nimmt auf dem Intervall  $[0, \pi/n]$  bei  $\pi/2n$  (und nur  
 dort) ihren maximalen Wert

$$\max_{\xi \in [0, 2\pi]} s_{[\psi]}(\xi) = \frac{1}{n} \sum_{j=0}^{n-1} \frac{1}{\sin(2j+1)\frac{\pi}{2n}} = \frac{2}{\pi} \ln n + O(1) \quad (4)$$

an.

Wir werden im 3. Abschnitt dieses Kapitels diese Lebesgue-  
 Konstante mit der Spanne einer i.a. nicht-optimalen  
 Funktionsstreifeneinschließung vergleichen.

### 3.1.2 Die diskrete Fouriertransformation

Zur Berechnung der Koeffizienten  $\alpha_j, \beta_j$  aus (2) kann man  
 folgendermaßen vorgehen (vgl. [50], S.66). Man setzt

$$w := \exp(-2\pi i/n) , \quad (5)$$

$$y_k := \eta_{2k} + i \eta_{2k+1} , \quad k=0(1)n-1 , \quad (6)$$

$$c_j := \frac{1}{n} \sum_{k=0}^{n-1} w^{jk} y_k , \quad j=0(1)n-1 , \quad (7)$$

und erhält  $\alpha_j, \beta_j$  aus

$$\left. \begin{aligned} \alpha_j - i\beta_j &= \frac{1}{2}(c_j + \bar{c}_{n-j}) + \frac{1}{2i}(c_j - \bar{c}_{n-j})\exp(-i\pi_j) , \\ \alpha_{n-j} - i\beta_{n-j} &= \frac{1}{2}(c_{n-j} + \bar{c}_j) + \frac{1}{2i}(\bar{c}_j - c_{n-j})\exp(i\pi_j) , \end{aligned} \right\} \quad (8)$$

für  $j = o(1)n$ , dabei ist  $c_n := c_0$  gesetzt.

(7) beschreibt eine lineare Transformation

$$c = w \eta \quad \text{mit } c = (c_k), \quad w := \frac{1}{n} (w^{jk}), \quad \eta = (\eta_k); \quad (9)$$

$c$  heißt die (diskrete) Fouriertransformierte zu  $\eta$ .

Wir wollen uns nun der Intervallinterpolationsaufgabe vom Typ I zuwenden. Aufgrund der Umformung (6) haben wir es jetzt mit "komplexen" Intervallen zu tun. Für diese werden wir zwei Arithmetiken, die sog. Rechteck- und Kreisscheibenarithmetik (vgl. [2], Abschnitte 5 und 6) im folgenden Abschnitt angeben. In Abschnitt 3 werden wir die Wertmenge und deren optimale Intervalleinschließung der Transformation (9) auf einem komplexen Intervallvektor bestimmen und anschließend eine Einschließung dieser Wertmenge bei einer speziellen Faktorisierung der Matrix  $w$  - dies ist gerade die FFT - untersuchen. In (8) treten auf der rechten Seite jeweils zwei Größen je zweimal auf. Aufgrund dieser Abhängigkeit erhält man bei der intervallmäßigen Durchführung obiger Transformation i.a. nicht mehr die optimalen Intervalleinschließungen  $[a_j], [b_j]$  aus (2').

### 3.2 Die Rechteck- und die Kreisscheibenarithmetik und Einschließungssätze für Produkte komplexer Intervallmatrizen

Mit kleinen deutschen Buchstaben bezeichnen wir komplexe Matrizen. Hinsichtlich der Ordnungsrelation

$$a \leq b : \Leftrightarrow \left\{ \begin{array}{l} \operatorname{Re} a_{jk} \leq \operatorname{Re} b_{jk}, \\ \operatorname{Im} a_{jk} \leq \operatorname{Im} b_{jk}, \end{array} \right\} j=1(1)r, k=1(1)s, \\ \text{für } a, b \in M_{r,s}(c)$$

ist der reelle Vektorraum  $M_{r,s}(C)$  ein bedingt vollständiger Vektorverband. Hinsichtlich dieser Ordnungsrelation definieren wir wie im Reellen für eine nichtleere, beschränkte Teilmenge  $\Theta$  des  $M_{r,s}(C)$

$$\text{Intv } \Theta := [\inf \Theta, \sup \Theta],$$

die optimale Intervalleinschließung von  $\Theta$ .

### 3.2.1 Rechteck- und Kreisscheibenarithmetik

Definition (Rechteck, Rechteckarithmetik, Spanne, Betrag):

Es seien  $[\alpha^{(1)}], [\alpha^{(2)}] \in II(R)$ , dann nennt man die Menge A

$$A := \{\alpha^{(1)} + i\alpha^{(2)} \mid \alpha^{(v)} \in [\alpha^{(v)}], v=1,2\}$$

Rechteck und schreibt  $[\alpha^{(1)}] + i[\alpha^{(2)}]$  für A.  $[\alpha^{(1)}]$  heißt der Realteil (ReA) und  $[\alpha^{(2)}]$  der Imaginärteil (ImA) von A. Diese Definitionen entsprechen der gängigen Sprechweise:

$A = [\alpha^{(1)}] + i[\alpha^{(2)}]$  stellt in der komplexen Ebene ein achsenparalleles Rechteck mit den Seiten  $[\alpha^{(1)}]$  und  $[\alpha^{(2)}]$  dar. Die Menge der Rechtecke sei mit  $R(C)$  bezeichnet. Die komplexen Zahlen können als einpunktige Rechtecke aufgefaßt werden; dann ist insbesondere  $II(R)$  in  $R(C)$  enthalten.

Für  $A = [\alpha^{(1)}] + i[\alpha^{(2)}]$ ,  $B = [\beta^{(1)}] + i[\beta^{(2)}] \in R(C)$  wird die Rechteckaddition, -subtraktion und -multiplikation definiert durch

$$\begin{aligned} A \pm B &:= [\alpha^{(1)}] \pm [\beta^{(1)}] + i[\alpha^{(2)}] \pm i[\beta^{(2)}] , \\ A * B &:= [\alpha^{(1)}][\beta^{(1)}] - [\alpha^{(2)}][\beta^{(2)}] \\ &\quad + i([\alpha^{(1)}][\beta^{(2)}] + [\alpha^{(2)}][\beta^{(1)}]) . \end{aligned}$$

Für  $A * B$  schreiben wir auch abkürzend  $AB$ .

Unter dem Konjugiert-komplexen Intervall  $\bar{A}$  zu

$A = [\alpha^{(1)}] + i[\alpha^{(2)}]$  verstehen wir

$$\bar{A} := [\alpha^{(1)}] - i[\alpha^{(2)}] .$$

Der Betrag und die Spanne eines Rechteckes

$A = [\alpha^{(1)}] + i[\alpha^{(2)}]$  werden definiert durch

$$\begin{aligned} |A| &:= |[\alpha^{(1)}]| + |[\alpha^{(2)}]| , \\ \text{span } A &:= \text{span}[\alpha^{(1)}] + \text{span}[\alpha^{(2)}] . \end{aligned}$$

Definition (Kreisscheibe, Kreisscheibenarithmetik,  
Betrag und Spanne):

Es sei  $a \in \mathbb{C}$  und  $\alpha \geq 0$ . Die Menge

$$\{z \in \mathbb{C} \mid |z-a| \leq \alpha\}$$

heißt Kreisscheibe mit Mittelpunkt  $a$  und Radius  $\alpha$  und wird mit  $\langle a, \alpha \rangle$  notiert (vgl. [2], S. 59/60). Die Menge der Kreisscheiben sei mit  $K(\mathbb{C})$  bezeichnet. Die komplexen Zahlen lassen sich als einpunktige Kreisscheiben (mit Radius 0) auffassen.

Verknüpfungen von Kreisscheiben werden wie folgt eingeführt:

Für  $A = \langle a, \alpha \rangle$ ,  $B = \langle b, \beta \rangle \in K(\mathbb{C})$  wird definiert:

$$A + B := \langle a+b, \alpha+\beta \rangle ,$$

$$A * B := \langle a*b, |a|\beta + |b|\alpha + \alpha\beta \rangle ;$$

insbesondere erhalten wir für  $c \in \mathbb{C}$

$$c * A = \langle ca, |c|\alpha \rangle .$$

Für  $A * B$  schreiben wir wieder abkürzend  $AB$ .

Den Betrag und die Spanne einer Kreisscheibe  $\langle a, \alpha \rangle$  definieren wir durch

$$\text{span} A := 2\alpha ,$$

$$|A| := |a| + \alpha .$$

Mit  $\text{II}(\mathbb{C})$  bezeichnen wir im folgenden, wenn nichts anderes vermerkt wird, entweder die Menge  $R(\mathbb{C})$  mit der Rechteckarithmetik oder die Menge  $K(\mathbb{C})$  mit der Kreisscheibenarithmetik. Eigenschaften von  $R(\mathbb{C})$  und  $K(\mathbb{C})$  entnehme man [2], Abschnitte 5 und 6. Es sei hier nur auf die für das folgende wichtigen Beziehungen hingewiesen:

Für  $A, B, C \in \text{II}(\mathbb{C})$  und  $a \in \mathbb{C}$  gilt

$$A(B+C) \subseteq AB + AC \quad (\text{Subdistributivität}) , \quad (10a)$$

$$a(B+C) = aB + aC \quad (\text{Semidistributivität}). \quad (10b)$$

Für  $A, B, C \in K(\mathbb{C})$  gilt das Assoziativgesetz bzgl. der Multiplikation

$$(AB)C = A(BC) .$$

Es sei noch ausdrücklich darauf hingewiesen, daß im Falle der Multiplikation in  $R(\mathbb{C})$  des Assoziativgesetz i.a. nicht gilt.

Es gilt jedoch

Lemma 3.1: Für  $a, b \in \mathbb{C}$  und  $C \in R(\mathbb{C})$  ist

$$(ab)C \subseteq a(bC) . \quad (10c)$$

Falls darüberhinaus  $a$  oder  $b$  reell oder rein imaginär ist, gilt sogar die Gleichheit.

Bemerkung: Falls  $\operatorname{Re} C = 0$  oder  $\operatorname{Im} C = 0$  ist, gilt i.a. keine Gleichheit in (10c).

Beweis: Es seien  $a = \rho + i\sigma$ ,  $b = \tau + i\omega$ ,  $C = [\gamma] + i[\delta]$ ; dann ist

$$\begin{aligned} (ab)C &= (\rho\tau - \sigma\omega + i(\rho\omega + \sigma\tau))([\gamma] + i[\delta]) \\ &= (\rho\tau - \sigma\omega)[\gamma] - (\rho\omega + \sigma\tau)[\delta] + i((\rho\tau - \sigma\omega)[\delta] + (\rho\omega + \sigma\tau)[\gamma]), \\ a(bC) &= (\rho + i\sigma)(\tau[\gamma] - \omega[\delta] + i(\tau[\delta] + \omega[\gamma])) \\ &= \rho(\tau[\gamma] - \omega[\delta]) - \sigma(\tau[\delta] + \omega[\gamma]) \\ &\quad + i(\rho(\tau[\delta] + \omega[\gamma]) + \sigma(\tau[\gamma] - \omega[\delta])), \end{aligned}$$

woraus sich die Behauptung ergibt. ▲

Wir werden dort, wo Assoziativität bzgl. der Multiplikation gilt, darauf verzichten, die Reihenfolge der Ausführung der Multiplikation kenntlich zu machen.

Wir weisen noch auf die im folgenden benötigten Regeln für das Rechnen mit Spanne und Betrag hin (vgl. [2], S. 72):

Lemma 3.2: Für  $A, B \in II(\mathbb{C})$  und  $a = \alpha^{(1)} + i\alpha^{(2)} \in \mathbb{C}$  gilt

$$\text{span}(A+B) = \text{span } A + \text{span } B,$$

$$\text{span}(aA) = \begin{cases} (|\alpha^{(1)}| + |\alpha^{(2)}|) \text{span } A & , \text{ falls } II(\mathbb{C}) = R(\mathbb{C}) \text{ ist,} \\ |a| \text{span } A & , \text{ falls } II(\mathbb{C}) = K(\mathbb{C}) \text{ ist,} \end{cases}$$

$$|A+B| \leq |A| + |B|,$$

$$|aA| = \begin{cases} (|\alpha^{(1)}| + |\alpha^{(2)}|) |A| & , \text{ falls } II(\mathbb{C}) = R(\mathbb{C}) \text{ ist,} \\ |a| |A| & , \text{ falls } II(\mathbb{C}) = K(\mathbb{C}) \text{ ist.} \end{cases}$$

### 3.2.2 Der Raum der komplexen Intervallmatrizen

Die Elemente von  $M_{r,s}(II(\mathbb{C}))$  heißen komplexe Intervallmatrizen, im Fall  $s=1$  auch komplexe Intervallvektoren, und werden mit großen deutschen Buchstaben bezeichnet:  $\mathfrak{U}, \mathfrak{B}, \mathfrak{C}, \dots$ . Wir identifizieren wieder die Punktmatrizen, d.h. die Matrizen, deren Komponenten sämtlich Punktintervalle sind, mit den Matrizen über  $\mathbb{C}$ . Die arithmetischen Operationen werden formal wie für die reellen Intervallmatrizen eingeführt (vgl. Abschnitt 1.2).

In Verallgemeinerung von Korollar 1.1 gilt:

Lemma 3.3: Für  $\mathfrak{U} \in M_{r,s}(R(\mathbb{C}))$ ,  $\mathfrak{B} \in M_{s,t}(R(\mathbb{C}))$  ist

$$\mathfrak{U}\mathfrak{B} = \text{Intv}\{a \cdot b \mid a \in \mathfrak{U}, b \in \mathfrak{B}\}.$$

Beweis: Für eine komplexe Intervallmatrix  $\mathfrak{C}$  bezeichne  $\text{Re } \mathfrak{C}$  bzw.  $\text{Im } \mathfrak{C}$  die (reelle) Intervallmatrix, die komponentenweise aus den Realteilen bzw. Imaginärteilen der entsprechenden Komponenten von  $\mathfrak{C}$  gebildet wird.

Es ist dann

$$\begin{pmatrix} \operatorname{Re} \mathfrak{U} & -\operatorname{Im} \mathfrak{U} \\ \operatorname{Im} \mathfrak{U} & \operatorname{Re} \mathfrak{U} \end{pmatrix} * \begin{pmatrix} \operatorname{Re} \mathfrak{B} \\ \operatorname{Im} \mathfrak{B} \end{pmatrix} = \begin{pmatrix} \operatorname{Re} (\mathfrak{U}\mathfrak{B}) \\ \operatorname{Im} (\mathfrak{U}\mathfrak{B}) \end{pmatrix} .$$

Die Aussage des Lemmas ergibt sich dann, indem Korollar 1.1 auf das derart modifizierte reelle Intervallproblem angewendet wird.  $\blacktriangle$

Lemma 3.4: Es seien  $\mathfrak{a} = (a_{ij})$ ,  $\mathfrak{b} = (b_{jk}) \in M_{s,s}(\mathbb{C})$  und  $\mathfrak{y} \in M_{s,1}(\operatorname{II}(\mathbb{C}))$ .

Dann gilt

$$(\mathfrak{a} \mathfrak{b}) \mathfrak{y} \subseteq \mathfrak{a} (\mathfrak{b} \mathfrak{y}) . \quad (11)$$

Es gelte darüberhinaus  $s := 2m$ ,

$$\left. \begin{aligned} \mathfrak{a} &:= \begin{pmatrix} \mathfrak{a}_1 & 0 \\ 0 & \mathfrak{a}_2 \end{pmatrix} \quad \text{mit den } m \times m\text{-Matrizen } \mathfrak{a}_1, \mathfrak{a}_2 \\ \text{und} \\ \mathfrak{b} &:= \begin{pmatrix} \mathfrak{c}_1 & \mathfrak{c}_2 \\ \mathfrak{c}_3 & \mathfrak{c}_4 \end{pmatrix} \quad \text{mit den } m \times m\text{-Diagonalmatrizen } \mathfrak{c}_v, v = 1, 2, 3, 4. \end{aligned} \right\} \quad (12)$$

Dann verschärft sich die Inklusion (11) für  $\mathfrak{y} \in M_{s,1}(K(\mathbb{C}))$  zur Mengengleichheit.

Beweis: Zum Beweis von beiden Teilen der Behauptung wird (Ob) und Lemma 3.1 angewendet; der Beweis soll hier nur für den ersten Teil der Behauptung erfolgen.

Es gilt für die  $j$ -te Komponente des Vektors  $(\mathfrak{a} \mathfrak{b}) \mathfrak{y}$

$$\begin{aligned} \sum_{k=1}^s \left( \sum_{u=1}^s a_{ju} b_{uk} \right) y_k &\subseteq \sum_{k=1}^s \sum_{u=1}^s (a_{ju} b_{uk}) y_k \\ &\subseteq \sum_{k=1}^s \sum_{u=1}^s a_{ju} (b_{uk} y_k) = \sum_{u=1}^s a_{ju} \sum_{k=1}^s b_{uk} y_k . \quad \blacktriangle \end{aligned}$$

### 3.3 Die intervallmäßige Durchführung der schnellen Fouriertransformation mit der Rechteck- und der Kreisscheibenarithmetik

Zur Behandlung der Intervallinterpolationsaufgabe setzen wir (6) entsprechend

$$Y_k := [\gamma_{2k}] + i[\gamma_{2k+1}], \quad k=0(1)n-1.$$

Gesucht ist nun

$$\{c\} := \{w \eta \mid \eta \in \mathcal{Y}\},$$

die Wertemenge der Transformation  $w$  (vgl. (9)) auf dem komplexen Intervallvektor  $\mathcal{Y} = (Y_k) \in M_{n,1}(R(\mathbb{C}))$ . Nach Lemma 3.3 gilt

$$\text{Intv } \{c\} = w \mathcal{Y}, \quad (13)$$

die optimale Intervalleinschließung der Menge  $\{c\}$  erhält man also durch Multiplikation der Punktmatrix  $w$  mit dem Intervallvektor  $\mathcal{Y}$ . Wie man sich leicht klar macht, geht die konvexe Hülle der Eckpunkte von  $\mathcal{Y}$  und damit auch  $\mathcal{Y}$  selbst unter der Abbildung  $w$  in die konvexe Hülle der Bilder der Eckpunkte von  $\mathcal{Y}$  über. Wir werden uns jedoch fortan auf die Bestimmung von  $\text{Intv } \{c\}$  gemäß (13) beschränken, da die Berechnung der konvexen Hülle der Bilder der Eckpunkte von  $\mathcal{Y}$  mit erheblichem Aufwand verbunden ist. Um nämlich zu garantieren, daß diese konvexe Hülle auch bei der praktischen Rechnung auf einem Computer eingeschlossen bleibt, ist die Transformation  $w$  für jeden der Eckpunkte mit einer Maschinenintervallarithmetik (vgl. Unterabsch. 3.4.1) durchzuführen und anschließend noch die konvexe Hülle der transformierten Vektoren zu bestimmen.

Bevor wir uns der intervallmäßigen Durchführung der FFT zuwenden, möchten wir an einem Beispiel zeigen, daß Algorithmen, die im Reellen oder Komplexen 'schnell' sind, für die Intervallrechnung unbrauchbar sein können. Wir betrachten dazu die 'schnelle' Multiplikation komplexer Zahlen (vgl. [36]).

Es seien

$$a = \alpha^{(1)} + i \alpha^{(2)}, b = \beta^{(1)} + i \beta^{(2)}. \quad (14)$$

Dann lassen sich Real- und Imaginärteil von  $c$ ,

$$c = \gamma^{(1)} + i \gamma^{(2)}, c = ab, \quad (15)$$

nach folgendem Algorithmus berechnen:

$$\left. \begin{aligned} \theta_1 &:= \gamma^{(1)} + \alpha^{(2)}, \theta_2 := \beta^{(1)} + \beta^{(2)}, \theta_3 := \beta^{(1)} - \beta^{(2)}, \\ \xi_1 &:= \theta_1 \beta^{(1)}, \quad \xi_2 := \theta_2 \alpha^{(2)}, \quad \xi_3 := \theta_3 \alpha^{(1)}, \\ \gamma^{(1)} &= \xi_1 - \xi_2, \quad \gamma^{(2)} = \xi_1 - \xi_3. \end{aligned} \right\} \quad (16)$$

Dieser Algorithmus benötigt drei reelle Multiplikationen und fünf reelle Additionen/Subtraktionen im Vergleich zu vier Multiplikationen und zwei Additionen/Subtraktionen bei der 'gewöhnlichen' Multiplikation komplexer Zahlen

$$\begin{aligned} \gamma^{(1)} &= \alpha^{(1)} \beta^{(1)} - \alpha^{(2)} \beta^{(2)}, \\ \gamma^{(2)} &= \alpha^{(1)} \beta^{(2)} + \alpha^{(2)} \beta^{(1)}. \end{aligned}$$

Da die Berechnung eines Produktes zweier reeller Intervalle wesentlich aufwendiger ist als die Berechnung ihrer Summe und ihrer Differenz (vgl. (1.1)), erscheint es sinnvoll, diesen Algorithmus intervallmäßig durchzuführen. Das bedeutet: Für  $A = [\alpha^{(1)}] + i[\alpha^{(2)}]$ ,  $B = [\beta^{(1)}] + i[\beta^{(2)}]$

werden in (16) alle reellen Größen durch reelle Intervalle und alle arithmetischen Operationen durch die entsprechenden intervallarithmetischen Operationen ersetzt und anschließend gesetzt

$$[\delta^{(1)}] := [\xi_1] - [\xi_2] ,$$

$$[\delta^{(2)}] := [\xi_1] - [\xi_3] .$$

Aufgrund der Inklusionsisotonie der intervallarithmetischen Operationen enthält dann offensichtlich das Rechteck

$$D = [\delta^{(1)}] + i[\delta^{(2)}] \text{ das Produkt } C := AB, C = [\gamma^{(1)}] + i[\gamma^{(2)}] .$$

Die Überschätzung von C durch D,  $\text{span } D - \text{span } C$ , kann jedoch (absolut und relativ) beliebig groß werden:

Setzt man beispielsweise

$$[\alpha^{(v)}] := \alpha^{(v)}[-1,1], [\beta^{(v)}] := \beta^{(v)}[-1,1] \\ \text{mit } \alpha^{(v)}, \beta^{(v)} \in \mathbb{R}_+, v = 1,2,$$

so erhält man mit Lemma 1.1 und den Setzungen (14) und (15)

$$\text{span}[\gamma^{(1)}] = 2(\alpha^{(1)}\beta^{(1)} + \alpha^{(2)}\beta^{(2)}) ,$$

$$\text{span}[\gamma^{(2)}] = 2\gamma^{(2)},$$

$$\text{span}[\delta^{(1)}] = 2(\alpha^{(1)}\beta^{(1)} + \alpha^{(2)}\beta^{(2)}) + 4\alpha^{(2)}\beta^{(1)},$$

$$\text{span}[\delta^{(2)}] = 2\gamma^{(2)} + 4\alpha^{(1)}\beta^{(1)} .$$

Damit ist

$$\text{span } D - \text{span } C = 4\beta^{(1)}(\alpha^{(1)} + \alpha^{(2)}) .$$

Wir wenden uns nun der FFT zu. Schwarz gibt in [49] eine sehr elementare und durchsichtige Darstellung der FFT an, die auf einer speziellen Faktorisierung der zeilenpermultierten Transformationsmatrix  $w$  in Blockmatrizen

$$\frac{1}{n} w_{p-1} \dots w_1 w_0 = p w, \quad (17)$$

( $p$  Permutationsmatrix)

beruht. Grundgedanke dieser Faktorisierung ist die Beobachtung, daß eine Fouriertransformation der Ordnung  $n=2^p$  auf zwei Fourier-Transformationen je der Ordnung  $m$  zurückgeführt werden kann.

Zur Definition der Matrizen  $w_0, w_1, \dots, w_{p-1}$  werden wir von der Binärdarstellung der Indizes Gebrauch machen. Wir nehmen dazu an, daß jede natürliche Zahl  $h \in \{0, 1, \dots, n-1\}$  binär dargestellt sei:

$$h = 2^{p-1}h_0 + 2^{p-2}h_1 + \dots + 2h_{p-2} + h_{p-1}$$

mit  $h_k \in \{0, 1\}$  für  $k = 0(1)p-1$ .

Wir definieren für  $k = 0(1)p-1$  die Matrix  $w_k =$

$$(w_k(h, j))_{h, j=0(1)n-1} \text{ durch } ^1) \quad (18)$$

$$w_k(h, j) := \begin{cases} \begin{cases} 1 & \text{falls } j = h, \\ 1 & \text{falls } j = h+2^{p-1-k}, \\ 0 & \text{sonst,} \end{cases} & \text{im Falle } h_k = 0, \\ \begin{cases} w_k(h) & \text{falls } j = h-2^{p-1-k} \\ -w_k(h) & \text{falls } j = h, \\ 0 & \text{sonst,} \end{cases} & \text{im Falle } h_k = 1, \end{cases}$$

---

1) Der Übersichtlichkeit halber schreiben wir fortan Indizes auch in Klammern hinter die zu indizierende Größe.

wobei  $w_k(h) := w^v$  mit

$$v := 2^k (2^{p-1-k} h_k + 2^{p-2-k} h_{k+1} + \dots + 2 h_{p-2} + h_{p-1}) - 2^{p-1} h_k \quad (19)$$

gesetzt ist.

Es gilt dann für  $\eta \in \mathbb{C}^n$  und  $h = o(1)_{n \rightarrow 1}$

$$((\frac{1}{n} w_{p-1} \dots w_0)(\eta))(h) = (p(w \eta))(h) = (w \eta)(h^*)$$

$$\text{mit } h^* := 2^{p-1} h_{p-1} + 2^{p-2} h_{p-2} + \dots + 2 h_1 + h_0.$$

Diese Umordnung der transformierten Größen wird Bitinversion genannt.

Wir setzen

$$\hat{w} := p w.$$

Die explizite Durchführung dieser sukzessiven Transformationen für die Fälle  $p = 2, 3, 4$  entnehme man [49].

Zur Behandlung der Intervallinterpolationsaufgabe sei nun wieder  $\eta \in M_{n,1}(R(\mathbb{C}))$  angenommen.

Wenn ein reelles (oder komplexes) Verfahren durch Übergang zu Intervall-Daten und -Operationen als ein Intervall-Verfahren geschrieben wird, muß i.a. mit einer Überschätzung gerechnet werden. Ein Beispiel hierfür ist die weiter oben diskutierte 'schnelle' Multiplikation komplexer Zahlen. So gilt auch hier

Satz 3.1: Für  $\eta \in M_{n,1}(R(\mathbb{C}))$  ist

$$\hat{w} \eta \subseteq \frac{1}{n} w_{p-1} (\dots (w_1 (w_0 \eta)) \dots).$$

Im Fall  $p \leq 3$  gilt die Gleichheit, für  $p > 3$  liefert die rechte Seite i.a. nur noch eine echte Obermenge von  $\hat{w} \eta$ .

Beweis: Die Inklusion ergibt sich durch Induktion aus Lemma 3.4 .

Der Beweis für die Behauptung im Fall  $p \leq 3$  soll hier nur für den Fall  $p = 3$  erfolgen.

Wir zeigen hier die Gültigkeit von

$$8(\hat{m} \eta)(h) = (m_2(m_1(m_3 \eta)))(h), \quad h = o(1)7, \quad$$

nur für den Fall  $h = 7$ . Für die anderen Komponenten verläuft der Beweis analog.

Es ist

$$(m_2(m_1(m_3 \eta)))(7) =$$

$$((Y(0)-Y(4))-(Y(1)-Y(6))w^2)-((Y(1)-Y(3))w-(Y(3)-Y(7))w^2)w^2)$$

Wegen  $w^2 = -i$  und nach Lemma 3.1 ergibt sich die rechte Seite zu  $8(\hat{m} \eta)(7) = 8(m \eta)(7)$ .

Zum Beweis der Behauptung im Fall  $p > 3$  geben wir ein Gegenbeispiel an. Es sei  $p = 4$ ,  $\eta \in M_{16,1}(R(C))$  mit  $Y(0) := 0$ ,  $Y(1) := [0,1]$ ,  $Y(h) := 0$ ,  $h = 2(1)15$ . Dann gilt

$$n(\hat{m} \eta)(15) = n(m \eta)(15) = iw^3 Y(1)$$

und

$$\begin{aligned} g &:= (m_3(m_2(m_1(m_4 \eta))))(15) \\ &= -w^4(w^2(wY(1))) = i(w^2(wY(1))). \end{aligned}$$

Es ist

$$w^2(wY(1)) = \frac{1}{2} \sqrt{2} (1-i) ((\cos \frac{\pi}{8} - i \sin \frac{\pi}{8}) [0,1])$$

mit dem Realteil

$$\frac{1}{2}\sqrt{2}(\cos \frac{\pi}{8}[0,1] - \sin \frac{\pi}{8}[0,1]) = \frac{1}{4}\sqrt{2}[-\sqrt{2-\sqrt{2}} , \sqrt{2+\sqrt{2}}]$$

$$\neq \frac{1}{4}\sqrt{2}(\sqrt{2+\sqrt{2}} - \sqrt{2-\sqrt{2}})[0,1] = \cos \frac{3\pi}{8}[0,1] ,$$

dem Realteil von  $w^3 Y(1)$  ; damit ist aber  $n(\hat{w} \eta)(15) \neq 8$   
und es folgt die Behauptung.  $\blacktriangle$

Daß aber im Gegensatz zur Aussage des Satzes 3.1 für  
Vektoren  $\eta \in M_{n,1}(K(c))$  bei der Faktorisierung (17)  
keine Überschätzung des Produktes  $\hat{w} \eta$  eintritt, ist die  
besonders erfreuliche Konsequenz des folgenden Satzes:

Satz 3.2: Es gilt für  $\eta \in M_{n,1}(K(c))$

$$\hat{w} \eta = \frac{1}{n} w_{p-1}(\dots(w_1(w_0 \eta))\dots) .$$

Beweis: Die Behauptung ist im Fall  $p = 1$  trivial erfüllt.

Im Fall  $p > 1$  genügt es zu zeigen, daß für alle  $\beta \in M_{n,1}(K(c))$   
und  $k = o(1)p-2$

$$(w_{p-1} \dots w_{k+1})w_k \beta = (w_{p-1} \dots w_k) \beta \quad (20)$$

gilt.

Wir definieren für  $k = o(1)p-1$ :  $r_k := w_{p-1} \dots w_k$  .

Nach (18) gilt für  $h, j = o(1)n-1$  und  $k = 1(1)p-1$

$$w_k(h, j) \neq 0 \Rightarrow h_{\kappa} = j_{\kappa} \text{ für } \kappa = o(1)\kappa-1 .$$

$w_k$  ist damit eine Blockdiagonalmatrix mit der Block-  
länge  $2^{p-k}$  , d.h.

$$w_k = \text{diag}(c_1, c_2, \dots, c_{2^k}), \text{ wobei } c_{11} \text{ } 2^{p-k} \times 2^{p-k}\text{-Matrix}$$

$$\text{für } 11 = 1(1)2^k \text{ ist.}$$

Damit ist auch das Produkt zweier Matrizen  $w_{k+\sigma}$ ,  $w_{k+\sigma}$ ,  $\sigma, \sigma = o(1)p-k-1$ , eine Blockdiagonalmatrix mit Blocklänge  $2^{p-k-\min\{\sigma, \sigma\}}$ . Insbesondere folgt hieraus, daß  $r_{k+1} \cdot w_k$  eine Blockdiagonalmatrix mit Blocklänge  $2^{p-k}$  ist. Im Fall  $k = 0$  sind  $r_1$  und  $w_0$  spezielle Matrizen der Gestalt  $a$  und  $b$  in (12), woraus sich für diesen Fall unmittelbar die Gültigkeit von (20) ergibt. Es sei fortan  $k \in \{1, 2, \dots, p-2\}$  gewählt.

Wir setzen noch

$$I_k(h) := \{j \in \{0, 1, \dots, n-1\} \mid h_{\kappa} = j_{\kappa}, \kappa = o(1)k-1\}.$$

Damit ist  $(r_{k+1} w_k)(h, j) = 0$ , falls  $j \notin I_k(h)$  ist.

Aus (18) folgt für  $j \in I_k(h)$

$$\begin{aligned} (r_{k+1} w_k)(h, j) &= \sum_{u=0}^{n-1} r_{k+1}(h, u) w_k(u, j) = \\ &= \begin{cases} r_{k+1}(h, j) w_k(j, j) + r_{k+1}(h, j+2^{p-1-k}) w_k(j+2^{p-1-k}, j) \\ \text{falls } j_k = 0, \\ r_{k+1}(h, j-2^{p-1-k}) w_k(j-2^{p-1-k}, j) + r_{k+1}(h, j) w_k(j, j), \\ \text{falls } j_k = 1, \end{cases} \\ &= \begin{cases} r_{k+1}(h, j) & , \text{falls } h_k = 0 \\ r_{k+1}(h, j+2^{p-1-k}) w_k(j+2^{p-1-k}) & , \text{falls } h_k = 1 \end{cases} \text{ und } j_k = 0, \\ &= \begin{cases} r_{k+1}(h, j-2^{p-1-k}) & , \text{falls } h_k = 0 \\ -r_{k+1}(h, j) w_k(j) & , \text{falls } h_k = 1 \end{cases} \text{ und } j_k = 1. \end{aligned}$$

Damit gilt für  $g \in M_{n,1}(K(c))$  und  $h = o(1)p-1$

$$\begin{aligned}
 ((r_{k+1}m_k)g)(h) &= \sum_{\substack{j \in I_k(h) \\ j_k=0}} (r_{k+1}m_k)(h, j) Z(j) \\
 &+ \sum_{\substack{j \in I_k(h) \\ j_k=1}} (r_{k+1}m_k)(h, j) Z(j) =
 \end{aligned}$$

$$\left\{ \begin{array}{l} \sum_{\substack{j \in I_k(h) \\ j_k=0}} r_{k+1}(h, j) Z(j) + \sum_{\substack{j \in I_k(h) \\ j_k=1}} r_{k+1}(h, j-2^{p-1-k}) Z(j), \\ \hspace{15em} \text{falls } h_k = 0, \\ \\ \sum_{\substack{j \in I_k(h) \\ j_k=0}} r_{k+1}(h, j+2^{p-1-k}) w_k(j+2^{p-1-k}) Z(j) - \\ \hspace{15em} \sum_{\substack{j \in I_k(h) \\ j_k=1}} r_{k+1}(h, j) w_k(j) Z(j), \text{ falls } h_k = 1. \end{array} \right.$$

Unter Verwendung von (10b) folgt daraus

$$\begin{aligned}
 ((r_{k+1}m_k)g)(h) &= \\
 \left\{ \begin{array}{l} \sum_{\substack{j \in I_k(h) \\ j_k=0}} (Z(j) + Z(j+2^{p-1-k})) r_{k+1}(h, j), \hspace{1em} \text{falls } h_k = 0, \\ \\ \sum_{\substack{j \in I_k(h) \\ j_k=1}} (Z(j-2^{p-1-k}) - Z(j)) w_k(j) r_{k+1}(h, j), \text{ falls } h_k = 1, \end{array} \right.
 \end{aligned}$$

$$= \begin{cases} \sum_{j \in I_{k+1}(n)} (Z(j) + Z(j+2^{p-1-k})) r_{k+1}(h, j), & \text{falls } h_k = 0, \\ \sum_{j \in I_{k+1}(h)} (Z(j-2^{p-1-k}) - Z(j)) w_k(j) r_{k+1}(h, j), & \text{falls } h_k = 1, \end{cases}$$

$$= (r_{k+1}(w_k g))(h),$$

denn es ist nach (18)

$$(w_k g)(j) = \begin{cases} Z(j) + Z(j+2^{p-1-k}) & , \text{ falls } j_k = 0, \\ (Z(j-2^{p-1-k}) - Z(j)) w_k(j) & , \text{ falls } j_k = 1. \end{cases}$$

Damit ist (20) gezeigt.  $\blacktriangle$

Wegen der Aussagen der Sätze 3.1 und 3.2 liegt es nahe, für  $n \geq 4$  die FFT mit Vektoren über  $K(\epsilon)$  anstelle mit Vektoren über  $R(\epsilon)$  durchzuführen. Bei der Behandlung der Intervallinterpolationsaufgabe ist jedoch i.a. der zu transformierende Vektor  $\eta = (Y(h))$  ein Element von  $M_{n,1}(R(\epsilon))$ . Überdies verlangt die intervallmäßige Durchführung der Transformation (8) Rechtecke. Wir schließen daher  $\eta$  in einen Vektor aus  $M_{n,1}(K(\epsilon))$  und dessen Fouriertransformierte in einen Vektor aus  $M_{n,1}(R(\epsilon))$  ein.

Im einzelnen gehen wir dazu folgendermaßen vor:

(a) Wir schließen jede der Komponenten

$$\begin{aligned} Y(h) &= [\eta^{(1)}(h)] + i[\eta^{(2)}(h)] \\ &= [\underline{\eta}^{(1)}(h), \overline{\eta}^{(1)}(h)] + i[\underline{\eta}^{(2)}(h), \overline{\eta}^{(2)}(h)] \end{aligned}$$

des Vektors  $\eta \in M_{n,1}(R(\mathbb{C}))$  in die jeweils flächenkleinste Kreisscheibe  $D(h) = \langle d(h), \delta(h) \rangle \in K(\mathbb{C})$  ein, d.h. wir setzen für  $h = o(1)n^{-1}$

$$\left. \begin{aligned} d(h) &:= \frac{1}{2}(\underline{\eta}^{(1)}(h) + \overline{\eta}^{(1)}(h), \underline{\eta}^{(2)}(h) + \overline{\eta}^{(2)}(h)), \\ \delta(h) &:= \frac{1}{2}(\text{span}^2[\eta^{(1)}(h)] + \text{span}^2[\eta^{(2)}(h)])^{1/2}. \end{aligned} \right\} \quad (21)$$

Dann gilt

$$Y(h) \subseteq D(h), \quad h = o(1)n^{-1}.$$

(b) Wir führen die Fouriertransformation mit dem Vektor  $\mathfrak{D} = (D(h))$  durch.

(c) Die Transformierte  $\widehat{\mathfrak{m}} \cdot \mathfrak{D} \in M_{n,1}(K(\mathbb{C}))$  wird komponentenweise in das jeweils flächenkleinste Rechteck eingeschlossen, d.h. wir setzen

$$C_j^* := \cap \{Z \in R(\mathbb{C}) \mid (\widehat{\mathfrak{m}} \mathfrak{D})(j) \subseteq Z\}, \quad j = o(1)n^{-1}.$$

Ist  $(\widehat{\mathfrak{m}} \mathfrak{D})(j) = \langle a_j, \alpha_j \rangle$ , so erhält man

$$C_j^* = a_j + \alpha_j \begin{bmatrix} -1, 1 \\ 1 \end{bmatrix}, \quad j = o(1)n^{-1}. \quad (22)$$

Wir möchten auf zwei leichte Nachteile der geschilderten Vorgehensweise hinweisen:

(a) Falls sich für ein  $h$   $\text{span}[\eta_{2h}]$  sehr von  $\text{span}[\eta_{2h+1}]$  unterscheidet, liefert die Einschließung (21) eine starke Überschätzung des Rechtecks  $Y(h)$ . In der Praxis, insbesondere bei dem in der Einleitung geschilderten Problem, kann jedoch meist von Intervallen  $[\eta_j]$  mit gleicher Spanne ausgegangen werden.

(3) Gegenüber der Durchführung der FFT in der Rechteckarithmetik wird die Programmierung etwas aufwendiger, und es wird mehr Rechenzeit benötigt. Für die Durchführung der Transformation mit der Rechteckarithmetik auf einer Rechenanlage kann das in [49] angegebene ALGOL-Programm leicht zu einem Triplex-ALGOL-Programm (vgl. [55]) abgeändert werden. Es sind dazu im wesentlichen nur die im Vereinbarungsteil auftretenden real-Variablen als triplex-Variablen zu deklarieren. Wird die Transformation mit der Kreisscheibenarithmetik 'schnell' durchgeführt, so ist zu beachten, daß in (18) nur skalare Multiplikationen auftreten, und daß diese Skalare überdies sämtlich vom Betrag 1 sind. Hierdurch vereinfacht sich die Programmierung der auftretenden Kreisscheibenmultiplikationen erheblich (s. (66) im Unterabschnitt 3.4.4).

Das Ergebnis der FFT, gleichgültig ob mit der Rechteck- oder mit der Kreisscheibenarithmetik nach dem oben geschilderten Verfahren berechnet, sei mit  $\mathfrak{C} = (C_j) \in M_{n,1}(R(\mathbb{C}))$  bezeichnet.

#### Die intervallmäßige Durchführung der Transformation (8)

In Analogie zu (8) definieren wir für  $j = o(1)n$

$$K_j := \frac{1}{2}(C_j + \bar{C}_{n-j}) + \frac{1}{2i}(C_j - \bar{C}_{n-j})\exp(-i\xi_j), \quad (23)$$

dabei ist  $C_n := C_0$  gesetzt.

Es gilt dann für  $j = o(1)n$

$$K_{n-j} = \frac{1}{2}(C_{n-j} + \bar{C}_j) + \frac{1}{2i}(\bar{C}_j - C_{n-j}) \exp(i \xi_j).$$

Ist  $K_j = [\alpha_j] + i[\beta_j]$ ,  $j = o(1)n$ , so liefern  $[\alpha_j], [\beta_j]$  einschließende Intervalle für die Koeffizienten  $\alpha_j, \beta_j$  von  $\varphi$  in (3)<sup>1)</sup>.

Mit  $C_j = [\zeta_j^{(1)}] + i[\zeta_j^{(2)}]$ ,  $j = o(1)n$ , ist

$$\begin{aligned} 2K_j = & [\zeta_j^{(1)}] + [\zeta_{n-j}^{(1)}] + i([\zeta_j^{(2)}] - [\zeta_{n-j}^{(2)}]) + \frac{1}{i}([\zeta_j^{(1)}] - [\zeta_{n-j}^{(1)}]) \\ & + i([\zeta_j^{(2)}] + [\zeta_{n-j}^{(2)}]))(\cos \xi_j - i \sin \xi_j) \end{aligned}$$

und mit

$$\left. \begin{aligned} [\kappa_j] &:= [\zeta_j^{(1)}] + [\zeta_{n-j}^{(1)}], \\ [\lambda_j] &:= [\zeta_j^{(2)}] - [\zeta_{n-j}^{(2)}], \\ [\mu_j] &:= ([\zeta_j^{(2)}] + [\zeta_{n-j}^{(2)}]) \cos \xi_j - ([\zeta_j^{(1)}] - [\zeta_{n-j}^{(1)}]) \sin \xi_j, \\ [\nu_j] &:= ([\zeta_j^{(2)}] + [\zeta_{n-j}^{(2)}]) \sin \xi_j + ([\zeta_j^{(1)}] - [\zeta_{n-j}^{(1)}]) \cos \xi_j, \end{aligned} \right\} \quad (24)$$

gilt

$$2K_j = [\kappa_j] + i[\lambda_j] + [\mu_j] - i[\nu_j]$$

und somit

$$\left. \begin{aligned} [\alpha_j] &= \frac{1}{2}([\kappa_j] + [\mu_j]), \\ [\beta_j] &= \frac{1}{2}([\nu_j] - [\lambda_j]), \end{aligned} \right\} \quad j = o(1)n. \quad (25)$$

---

1) Es wird natürlich  $[\kappa_n] := [\beta_n] := 0$  gesetzt.  
Darauf wird im nachfolgenden nicht mehr ausdrücklich hingewiesen werden.

Analog erhält man

$$\left. \begin{aligned} \lceil \alpha_{n-j} \rceil &= \frac{1}{2}(\lceil x_j \rceil - \lceil u_j \rceil), \\ \lceil \beta_{n-j} \rceil &= \frac{1}{2}(\lceil x_j \rceil + \lceil v_j \rceil), \end{aligned} \right\} \quad j = o(1)n. \quad (26)$$

Damit erhalten wir die (i.a. nichtoptimale) Funktionsstreifeneinschließung

$$|\varphi|(\xi) := \frac{\lceil \alpha_n \rceil}{2} + \sum_{j=1}^{n-1} \lceil \alpha_j \rceil \cos j\xi + \lceil \beta_j \rceil \sin j\xi + \frac{\lceil \alpha_n \rceil}{2} \cos n\xi \quad (27)$$

für jedes trigonometrische Polynom, welches (3) erfüllt.

Wir wollen nun die Spanne dieser Einschließung abschätzen.

Dazu bezeichne der Intervallvektor  $\mathfrak{Y}_k = (Y_k(j))_{j=0}^{n-1}$ ,

$k = o(1)p$ , die Transformierte im  $k$ -ten Schritt, d.h.

$$\left. \begin{aligned} \mathfrak{Y}_0 &:= \frac{1}{n} \mathfrak{Y} \\ \mathfrak{Y}_{k+1} &:= \mathfrak{Y}_k \cdot \mathfrak{A}_k, \quad \mathfrak{A}_k = o(1)p^{-1}. \end{aligned} \right\} \quad (28)$$

Lemma 3.5: Zu  $k = o(1)p-1$ ,  $j=o(1)n-1$ , gibt es  $\varphi_{k,j,u}^{(k)}, \dots$ ,

$\varphi_{k,j,u}^{(o)} \in \mathbb{C}$ ,  $u = o(1)2^{k+1}-1$ , sämtlich vom Betrag 1, derart daß

$$\varphi_{k+1}^{(j)} = \sum_{i=0}^{2^{k+1}-1} \varphi_{k,j,u}^{(k)} (\dots (\varphi_{k,j,u}^{(o)} \varphi_o(\tau_k(j) + i \cdot 2^{p-k-1})) \dots), \quad (29)$$

wobei

$$\varphi_k^{(j)} := \begin{cases} \sum_{\alpha=0}^{2^{p-k-1}-1} j_{k+\alpha} 2^{p-k-\alpha-1} & \text{für } k \in \{0,1,\dots,p-2\}, \\ 0 & \text{für } k = p-1, \end{cases} \quad (1)$$

<sup>1)</sup> Wie üblich werde  $\dots := 0$  für  $n \leq 0$  gesetzt.

Für  $k = p - 1$  sind  $\varphi_{k,j,u}^{(k)}$  und  $\varphi_{k,j,u}^{(k-1)}$ ,  $j = 0(1)n-1$ ,  
 $u = 0(1)n-1$ , sämtlich reell oder rein imaginär.

Insbesondere folgt daraus, daß in  $Y_{k+1}(j)$  jede der komplexen Intervalle  $Y(h)$ ,  $h = 0(1)n-1$ , höchstens einmal auftritt.

Der Beweis erfolgt durch Induktion über  $k = 0(1)p-1$ :

Den Induktionsanfang folgert man leicht aus (18) für  $k = 0$ .

Annahme: Die Behauptung sei richtig für  $k \in \{0, 1, \dots, p-2\}$ .

Zu zeigen ist:

für  $j = 0(1)n-1$  gibt es  $\varphi_{k+1,j,u}^{(k+1)}, \dots, \varphi_{k+1,j,u}^{(0)} \in \mathbb{C}$ ,

$u = 0(1)2^{k+2}-1$ , sämtlich von Betrag 1, so daß

$$Y_{k+2}(j) = \sum_{u=0}^{2^{k+2}-1} \varphi_{k+1,j,u}^{(k+1)} (\dots (\varphi_{k+1,j,u}^{(0)} Y_0(\tau_{k+1}(j) + u2^{p-k-2})) \dots) \quad (30)$$

ist.

O.B.d.A. sei  $j_{k+1} = 1$ , dann gilt mit  $j' := j - 2^{p-k-2}$  nach (18)

$$\begin{aligned} Y_{k+2}(j) &= (Y_{k+1}(j') - Y_{k+1}(j)) w_{k+1}(j) \\ &= w_{k+1}(j) \sum_{u=0}^{2^{k+1}-1} \varphi_{k,j',u}^{(k)} (\dots (\varphi_{k,j',u}^{(0)} Y_0(\tau_k(j') + u2^{p-k-1})) \dots) \\ &\quad - w_{k+1}(j) \sum_{u=0}^{2^{k+1}-1} \varphi_{k,j,u}^{(k)} (\dots (\varphi_{k,j,u}^{(0)} Y_0(\tau_k(j) + u2^{p-k-1})) \dots) . \end{aligned}$$

Nun ist  $j_{k+1} = 1$ , und wir erhalten

$$\tau_k(j') = \sum_{i=0}^{p-k-1} j_{k+i} 2^{p-k-1-i} = \sum_{i=0}^{p-k-2} j_{k+i+1} 2^{p-k-2-i} = \tau_{k+1}(j) ,$$

$$\tau_k(j) = \tau_{k+1}(j) + 2^{p-k-2} .$$

Wegen

$$\begin{aligned} & \{u 2^{p-k-1}, u 2^{p-k-1} + 2^{p-k-2} | u = o(1) 2^{k+1-1}\} \\ &= \{u 2^{p-k-2} | u = o(1) 2^{k+2-1}\} \end{aligned}$$

folgt (30) mit

$$\varphi_{k+1,j,u}^{(\nu)} := \begin{cases} \varphi_{k,j',u}^{(\kappa)} , & \text{falls } u = 2\nu , \\ \varphi_{k,j,u}^{(\kappa)} , & \text{falls } u = 2\nu+1 , \end{cases}$$

für  $\kappa = o(1)k$  ,

$$\varphi_{k+1,j,u}^{(k+1)} := \begin{cases} w_{k+1}(j) , & \text{falls } u = 2\nu , \\ -w_{k+1}(j) , & \text{falls } u = 2\nu+1 , \end{cases}$$

für  $u = o(1) 2^{k+2-1}$ ,  $j = o(1)n-1$  .

Der Zusatz für  $k = p-1$  folgt aus (18) und (19).  $\blacktriangle$

Satz 3.3 : Es gelte

$$\text{span}[\eta_k] \leq \alpha , \quad k = o(1)2n-1 . \quad (31)$$

Wir setzen  $\alpha := 1 + \sqrt{2}$  und

$$x := \begin{Bmatrix} \sqrt{2} \\ 1 \\ \alpha/2 \\ \alpha/2 \end{Bmatrix} \text{ im Fall der } \begin{cases} \text{Kreisscheibenarithmetik} \\ \text{Rechteckarithmetik und } p \begin{cases} =1,2 \\ =3 \\ \geq 4. \end{cases} \end{cases}$$

Für die unter Verwendung der FFT berechneten Intervalle  $[\alpha_j]$ ,  $[\beta_j]$  gilt dann

$$\text{span}[\alpha_j], \text{span}[\beta_j] \leq \chi \cup \alpha, \quad (32)$$

und es ist für  $\xi \in [0, 2\pi[$  und den Funktionsstreifen (27)

$$\text{span}[\varphi](\xi) \leq \chi \cup \alpha (1 + (n-1)^{1/2}). \quad (33)$$

Bemerkung: 1.) Die Spanne einer Kreisscheibe ist nicht unmittelbar mit der Spanne eines Rechtecks vergleichbar. Ist  $A = [\alpha] + i[\alpha]$  ein nichtentartetes Quadrat und  $D$  die nach (21) definierte Kreisscheibeneinschließung zu  $A$ , dann gilt:

$$\text{span} A = 2\text{span}[\alpha] > \sqrt{2} \text{span}[\alpha] = \text{span} D,$$

obwohl  $A \subset D$  ist.

2.) In (4) hatten wir ein Anwachsen der Spanne der optimalen Funktionsstreifeneinschließung wie etwa  $\frac{2}{\pi} \ln n$  festgestellt. Hier wächst nun die obere Schranke für die Spanne der Funktionsstreifeneinschließung  $[\varphi]$  proportional zu  $n$  bzw.  $n^{3/2}$ .

Beweis: Es werden die Rechenregeln für die Spanne aus Lemma 1.1 und Lemma 3.2 verwendet.

Es wird zunächst die Aussage für die Kreisscheibenarithmetik gezeigt:

Für die Kreisscheibeneinschließung  $\mathfrak{T} = (D(j))$  aus (21) gilt unter der Voraussetzung (31)

$$\text{span} D(j) \leq \alpha^{1/2}, \quad j = o(1)n^{-1}.$$

Nach Satz 3.2 genügt es, die Spanne der Komponenten des Vektors  $m \mathcal{D}$  abzuschätzen, um eine obere Schranke für die Spanne der Komponenten des Vektors  $\frac{1}{n} m_{p-1}(\dots(m \mathcal{D})\dots)$  zu erhalten.

$$\begin{aligned} \text{span}(m \mathcal{D})(j) &= \frac{1}{n} \text{span}\left(\sum_{k=0}^{n-1} w^{jk} D(k)\right) \\ &= \frac{1}{n} \sum_{k=0}^{n-1} |w^{jk}| \text{span } D(k) \\ &= \frac{1}{n} \sum_{k=0}^{n-1} \text{span } D(k) \leq \alpha \sqrt{2} . \end{aligned}$$

Damit ergibt sich für die Rechteckeinschließung  $C_j =$

$$[\zeta_j^{(1)}] + i[\zeta_j^{(2)}] \text{ aus (22)}$$

$$\text{span}[\zeta_j^{(v)}] \leq \alpha \sqrt{2} , \quad v = 1, 2 . \quad (34)$$

Aus (24) und (25) folgt unter Verwendung von

$$|\cos \xi| + |\sin \xi| \leq \sqrt{2} . \quad (35)$$

mit (34)

$$\text{span}[\alpha_j], \text{span}[\beta_j] \leq \alpha \sqrt{2} (1 + |\cos \xi| + |\sin \xi|)$$

und damit (32).

Schließlich ist

$$\text{span}[\varphi](\varepsilon) \leq \alpha(2 + \sqrt{2}) + \sum_{j=1}^{n-1} (|\cos \xi| + |\sin \xi|) \alpha(2 + \sqrt{2}) ,$$

woraus (33) folgt.

Es werde nun die Durchführung der FFT mit der Rechteckarithmetik betrachtet. Die Aussagen für  $p \leq 3$  ergeben sich leicht durch Nachrechnen. Es sei nun  $p > 3$ . Jeder der Summanden der Summe in (29) besteht für  $k = p-1$  aus einer Komponente des Vektors  $\eta_k$ , die noch mit  $p$  komplexen Zahlen multipliziert ist. Von diesen komplexen Skalaren sind mindestens zwei reell oder rein imaginär. Mit Induktion zeigt man leicht, daß dann die Spanne jeweils des Real- und des Imaginärteils eines Summanden kleiner gleich  $(\sqrt{2})^{p-2} \alpha/n$  ist. Damit ist die Spanne des Real- bzw. Imaginärteils der Summe in (29) höchstens  $\sqrt{n} \alpha/2$ . Die restlichen Abschätzungen ergeben sich entsprechend denjenigen bei der Kreisscheibenarithmetik. ▲

Im Anhang geben wir ein Programm in Triplex-ALGOL an, welches Schranken für die Koeffizienten des trigonometrischen Interpolationspolynoms berechnet. Die FFT wird darin mit der Kreisscheibenarithmetik ausgeführt.

Im Unterabschnitt 4.2.1 werden wir eine Anwendung der intervallmäßigen Durchführung der FFT angeben.

### 3.4. Akkumulation der Rundungsfehler bei der intervallmäßigen Durchführung der schnellen Fouriertransformation und bei der Festimmung von Schranken für die Koeffizienten des trigonometrischen Interpolationspolynoms

In vielen Arbeiten (s. beispielsweise [25], [42] ist daraufhingewiesen worden, daß die FFT rundungsfehlerstabil ist.

Wir möchten uns nun der Frage zuwenden, mit welchem Anwachsen der Rundungsfehler bei der intervallmäßigen Durchführung der FFT und bei der Bestimmung der Intervalle  $[\alpha_j]$ ,  $[\beta_j]$  im ungünstigsten Falle gerechnet werden muß.

Im vorliegenden Abschnitt gehen wir zunächst kurz auf die Realisierung der Intervallarithmetik auf einem Computer ein. Eine obere Schranke für die Akkumulation der Rundungsfehler bei der Bestimmung der Intervalle  $[\alpha'_j]$ ,  $[\beta'_j]$  aus (2') leiten wir in Unterabschnitt 3.4.2 her. Anschließend behandeln wir in Unterabschnitt 3.4.3 die Akkumulation der Rundungsfehler bei der Durchführung der FFT in der Rechteckarithmetik. Das Hauptergebnis dieser Untersuchung ist in Satz 3.5 zusammengefaßt. Eine obere Schranke für die Rundungsfehler der gesamten Rechnung zur Bestimmung von  $[\alpha_j]$ ,  $[\beta_j]$ , d.h. für die FFT und die Transformationen (24), (25) bzw. (26), ist in Satz 3.6 angegeben. Im Anschluß daran zeigen wir, daß die obere Schranke für die Akkumulation der Rundungsfehler bei der intervallmäßigen Durchführung der FFT in der Kreisscheibenarithmetik günstiger ausfällt als diejenige für die Rechteckarithmetik. Derartige Rundungsfehlerschranken pflegen oft sehr pessimistisch zu sein und geben selten die Größe des tatsächlichen Fehlers an. Dies zeigen die numerischen Beispiele am Ende des vorliegenden Abschnitts. Für die Kreisscheibenarithmetik ist jedoch nicht nur die a-priori-Schranke für die Rundungsfehler kleiner als bei der Rechteckarithmetik, sondern auch die in praxi auftretende Rundungsfehlerakkumulation für  $n \geq 32$ .

Der Einfachheit halber werde im Fall  $e$  der FFT die Division durch  $n$  zu Beginn der Transformation wie in (28) vorgenommen und die Rundungsfehler, die bei dieser Division entstehen, nicht berücksichtigt. Entsprechendes gilt auch für die direkte Rechnung nach (2').

Die praktischen Rechnungen des vorliegenden Abschnitts wurden auf der UNIVAC 1100/81 (mit 8 Ziffern) des Rechenzentrums der Universität Freiburg i.Br. vorgenommen.

Wir schreiben fortan abkürzend auch  $s$  (...) für  $\text{span}$ .

#### 3.4.1 Maschinenintervallarithmetik

Wir bezeichnen die Menge der auf einem (für das Folgende festgewählten) Computer vorhandenen (normalisierten) Gleitkommazahlen mit  $\mathbb{R}_M$ . Es läßt sich dann mit den beiden (festen) natürlichen Zahlen  $l, b$  mit  $l \geq 1$  und  $b \geq 2$  jedes  $\tilde{a} \in \mathbb{R}_M^{1)}$  mit  $\tilde{a} \neq 0$  darstellen als

$$\tilde{a} = m\beta^e \quad \text{mit} \quad \beta^{-1} \leq m < 1 \quad \text{und} \quad e \in \mathbb{Z},$$

wobei

$$m = \sum_{v=1}^l a_v \beta^{-v} \quad \text{mit} \quad a_v \in \{0, 1, \dots, b-1\}$$

und  $a_1 \neq 0$  ist.

---

1) Um die Übersichtlichkeit in den Rechnungen der nachfolgenden Unterabschnitte zu wahren, notieren wir in diesem Abschnitt reelle Zahlen auch mit kleinen, reelle Intervalle auch mit großen lateinischen Buchstaben. Die Tilde soll andeuten, daß es sich hier um Maschinenzahlen und -intervalle sowie um Maschinen-Operationen zwischen diesen handelt.

Es werden  $m$  die Mantisse,  $B$  die Basis,  $e$  der Exponent und  $l$  die Mantissenlänge genannt. Für den Exponenten sollen der Einfachheit halber alle ganzen Zahlen zugelassen werden, Bereichsüber- oder -unterschreitungen also ausgeschlossen sein. Gewöhnlich ist die Basis  $B = 2$  oder  $B = 10$ . Die Zahl  $\tilde{0} \in \mathbb{R}_M$  wird gesondert dargestellt, etwa mit  $m = 0$  und  $e = 0$ .

Auf dem Computer seien zwei gerichtete Rundungen

$\downarrow, \uparrow : \mathbb{R} \rightarrow \mathbb{R}_M$ , eine Ab- und eine Aufrundung, vorhanden <sup>1)</sup>, d.h. für die Abbildungen  $\downarrow, \uparrow$  gelte

$$\forall a \in \mathbb{R} \quad \downarrow a \leq a \leq \uparrow a.$$

Ferner mögen die beiden gerichteten Rundungen die beiden folgenden Eigenschaften besitzen

$$\begin{aligned} \forall a, b \in \mathbb{R} : a \leq b &\Rightarrow (\downarrow a \leq \downarrow b \text{ und } \uparrow a \leq \uparrow b), \\ \forall a \in \mathbb{R}_M : \downarrow a &= \uparrow a = a. \end{aligned}$$

Unter einem Maschinenintervall  $[\underline{a}] = [\underline{a}, \bar{a}]$  verstehen wir ein Intervall aus Maschinenzahlen

$$[\underline{a}] := \{ \tilde{a} \in \mathbb{R}_M \mid \underline{a} \leq \tilde{a} \leq \bar{a} \} \quad \text{mit } \underline{a}, \bar{a} \in \mathbb{R}_M.$$

Mit  $II(\mathbb{R}_M)$  sei die Menge der Maschinenintervalle bezeichnet.

Definiert man eine Intervallrundung  $\updownarrow : II(\mathbb{R}) \rightarrow II(\mathbb{R}_M)$  gemäß

$$\forall [a] = [\underline{a}, \bar{a}] \in II(\mathbb{R}) : \updownarrow [a] := [\downarrow \underline{a}, \uparrow \bar{a}], \quad (36)$$

1) Liegt  $\mathbb{R}_M$  symmetrisch zur Null, d.h. gilt  $\mathbb{R}_M = -\mathbb{R}_M$ , dann genügt schon die Existenz einer gerichteten Rundung aus; ist nämlich beispielsweise eine Abrundung  $\downarrow$  vorhanden, dann erhält man eine Aufrundung  $\uparrow$  aus

$$\forall a \in \mathbb{R} : \uparrow a := -(\downarrow (-a))$$

(vgl. [2], S. 48).

dann gilt

$$\begin{aligned} \forall [a] \in II(\mathbb{R}): \quad [a] &\subseteq \uparrow [a] , \\ \forall [a], [b] \in II(\mathbb{R}): [a] &\subseteq [b] \rightarrow \uparrow [a] \subseteq \uparrow [b] . \end{aligned}$$

Wir wollen nun annehmen, daß die Basis  $B = 10$  ist, also ein Dezimalsystem vorhanden ist. Für die gerichteten Rundungen möge mit

$$\epsilon := 10^{1-1} \quad (37)$$

gelten

$$\forall a \in \mathbb{R} : |a - \downarrow a|, |a - \uparrow a| \leq \epsilon |a| . \quad (38)$$

Diese Bedingungen sind erfüllt, wenn zu der nächst oberen bzw. zu der nächst unteren Maschinenzahl gerundet wird (vgl. [50], S. 4-5).

Wir setzen

$$[\epsilon] := \epsilon [-1, 1] .$$

Für die oben definierte Intervallrundung  $\downarrow$  gilt dann

$$\forall [a] \in II(\mathbb{R}) : \downarrow [a] \subseteq [a] (1 + [\epsilon]) . \quad (39)$$

Jeder Operation  $w \in \Omega := \{+, -, *, /\}$  auf  $II(\mathbb{R}) \times II(\mathbb{R})$  ist eine Maschinenintervalloperation  $\tilde{w} \in \tilde{\Omega} := \{\tilde{+}, \tilde{-}, \tilde{*}, \tilde{/}\}$

$$\tilde{w} : II(\mathbb{Q}_M) \vee II(\mathbb{P}_M) \rightarrow II(\mathbb{Q}_M)$$

gemäß

$$\forall [\tilde{a}], [\tilde{b}] \in II(\mathbb{Q}_M) : [\tilde{a}] \tilde{w} [\tilde{b}] := \downarrow ([\tilde{a}] w [\tilde{b}])$$

zugeordnet. Mit (39) folgt daraus sofort

$$\left. \begin{aligned} [\tilde{\alpha}] \tilde{\omega} [\tilde{\beta}] &\subseteq ([\tilde{\alpha}] \omega [\tilde{\beta}]) (1 + [\epsilon]) \\ \text{für alle } [\tilde{\alpha}], [\tilde{\beta}] &\in II(R_M) \text{ und } \tilde{\omega} \in \tilde{\Omega} . \end{aligned} \right\} \quad (40)$$

Es folgt

Lemma 3.6 (vgl. etwa [2], S. 50/51): Es sei  $\uparrow$  eine Intervallrundung, die gemäß (2) von den gerichteten Rundungen  $\downarrow, \uparrow$  der Schranken erzeugt wird, und es sei  $\tilde{\omega} \in \tilde{\Omega}$ . Dann gilt

$$[\tilde{\alpha}] \omega [\tilde{\beta}] \subseteq [\tilde{\alpha}] \tilde{\omega} [\tilde{\beta}] \text{ für alle } [\tilde{\alpha}], [\tilde{\beta}] \in II(R_M) .$$

Eine Maschinenintervallarithmetik wurde beispielsweise mit Triplex-ALGOL 60 realisiert (s. [55]).

Mit  $\widetilde{\cos \xi_j}$  bzw.  $\widetilde{\sin \xi_j}$  seien Maschineneinschließungen für den Wert des Cosinus bzw. Sinus an der Stelle  $\xi_j$ ,  $j=0(1)2n-1$ , gegeben<sup>1)</sup>. Es gibt dann ein  $\delta_n \in R_+$  mit

$$\left. \begin{aligned} \widetilde{\cos \xi_j} &\subseteq \cos \xi_j (1 + \delta_n [\epsilon]) , \\ \widetilde{\sin \xi_j} &\subseteq \sin \xi_j (1 + \delta_n [\epsilon]) , \end{aligned} \right\} \quad j=0(1)2n-1 . \quad (41)$$

Die Konstante  $\delta_n$  ist von  $n$  abhängig, wenn die Einschließungen  $\cos \xi_j$ ,  $\sin \xi_j$  rekursiv berechnet werden. Wir werden uns in Unterabschnitt 3.4.5 mit der Bestimmung der Konstanten  $\delta_n$  befassen.

---

1) Im Fall  $j = 0$  gilt natürlich

$$\cos \xi_0 = 1 \quad \text{und} \quad \sin \xi_0 = 0 .$$

Diese Einschließungen liefern für

$$\left. \begin{aligned} u_k(j) &:= \operatorname{Re} w_k(j), \\ v_k(j) &:= \operatorname{Im} w_k(j), \end{aligned} \right\} \quad j=o(1)n-1, \quad k=o(1)p, \quad (42)$$

Einschließungen  $\tilde{U}_k(j), \tilde{V}_k(j)$  mit

$$\left. \begin{aligned} \tilde{U}_k(j) &\subseteq u_k(j) (1 + \delta_n[\epsilon]), \\ \tilde{V}_k(j) &\subseteq v_k(j) (1 + \delta_n[\epsilon]), \end{aligned} \right\} \quad j=o(1)n-1, \quad k=o(1)p. \quad (43)$$

In den nachfolgenden Unterabschnitten werden wir nach  $\epsilon$  linearisieren.

### 3.4.2 Direkte Berechnung

Um die Rundungsfehler für die direkte Rechnung nach (2') abzuschätzen, genügt es, sich auf die Bestimmung von  $[\alpha_j']$  zu beschränken.

Bei der Rechnung mit der in Unterabschnitt 3.4.1 eingeführten Maschinenintervallarithmetik erhält man mit

$$\left. \begin{aligned} [\theta_k] &:= \frac{1}{n} [\eta_k], \\ [\vartheta_k] &:= \uparrow [\theta_k], \end{aligned} \right\} \quad k=o(1)2n-1, \quad (44)$$

die Maschinenintervalleinschließung für  $[\alpha_j']$

$$[\tilde{\alpha}_j'] := \sum_{k=0}^{\widetilde{2n-1}} [\tilde{\vartheta}_k] * \widetilde{\cos j \xi_k}, \quad j=o(1)n. \quad (45)$$

Satz 3.4: Für die vorgegebene Maschinenintervallarithmetik gelte (37), (38) und (41). Dann läßt sich die Überschätzung der Einschließung  $[\alpha_j']$  durch die mit dieser Maschinenintervallarithmetik gemäß (45) berechneten Einschließung  $[\tilde{\alpha}_j']$  abschätzen durch

$$s([\tilde{\alpha}_j']) - s([\alpha_j']) \leq \frac{2}{n} \epsilon \cdot n (2n^2 + 5n - 2 + (2n-1)\delta_n) + O(\epsilon^2)$$

mit

$$n := \max\{|\eta_k| \mid k=0(1)2n-1\} . \quad (46)$$

Beweis: Aus (44) ergibt sich mit (39), (40) und (41) <sup>1)</sup>

$$\begin{aligned} [\tilde{\alpha}_j'] &\subseteq \bigcup_{k=0}^{2n-1} [\theta_k] \cos j \xi_k \quad (1+\delta_n[\epsilon])^{\text{sign } k} (1+[\epsilon])^{1+\text{sign } k} \\ &\subseteq [\theta_0] (1+[\epsilon])^{2n} \\ &\quad + \sum_{k=1}^{2n-1} [\theta_k] \cos j \xi_k \quad (1+\delta_n[\epsilon])(1+[\epsilon])^{2n-k+2} \\ &\subseteq [\theta_0] + [\theta_0] \sum_{u=0}^{2n-1} \binom{2n}{u} [\epsilon]^{2n-u} \\ &\quad + \sum_{k=1}^{2n-1} [\theta_k] \cos j \xi_k \\ &\quad + \sum_{k=1}^{2n-1} ([\theta_k] \cos j \xi_k \sum_{u=0}^{2n-k+1} \binom{2n-k+2}{u} [\epsilon]^{2n-k+2-u}) \\ &\quad + \sum_{k=1}^{2n-1} ([\theta_k] \cos j \xi_k \delta_n \sum_{u=0}^{2n-k+2} \binom{2n-k+2}{u} [\epsilon]^{2n-k+3-u}). \end{aligned}$$

---

1) Es ist  $[a]^0 := 1$  für  $|a| \in \text{II}(\mathbb{R})$  gesetzt.

Die Summanden  $[e]^\nu$  sind für  $\nu \geq 1$  symmetrisch zur Null;  
da die Summe zweier zu Null symmetrischer Intervalle  
wieder zur Null symmetrisch ist, folgt mit (1.2 d)

$$\begin{aligned} s([\tilde{\alpha}_j']) &\leq s([\alpha_0]) + 4n \epsilon |[\alpha_0]| \\ &\quad + \sum_{k=1}^{2n-1} s([\alpha_k]) |\cos j \xi_k| \\ &\quad + 2 \sum_{k=1}^{2n-1} |[\alpha_k]| |\cos j \xi_k| (2n-k+2+\delta_n) \epsilon + O(\epsilon^2). \end{aligned}$$

Mit (46) folgt hieraus

$$s([\tilde{\alpha}_j']) - s([\alpha_j']) \leq 4 \epsilon n + \frac{2}{n} \epsilon n \sum_{k=1}^{2n-1} (2n-k+2+\delta_n) + O(\epsilon^2),$$

woraus sich die verlangte Abschätzung ergibt. ▲

### 3.4.3 Schnelle Fouriertransformation mit der Rechteckarithmetik

Wir verwenden wieder die Bezeichnungen (28).

Nach (18) ist dann für  $k=o(1)p-1$  und  $j=o(1)n-1$

$$Y_{k+1}(j) = \left\{ \begin{array}{ll} Y_k(j) + Y_k(j+2^{p-1-k}) & , \text{ falls } j_k=0, \\ (Y_k(j-2^{p-1-k}) - Y_k(j))w_k(j), & \text{ falls } j_k=1, \end{array} \right\} \quad (47)$$

wobei  $w_k(j)$  nach (19) gegeben ist.

Real- und Imaginärteil von  $Y_k(j)$  können getrennt berechnet  
werden: Mit

$$\left. \begin{aligned} G_k(j) &:= \operatorname{Re} Y_k(j), \\ H_k(j) &:= \operatorname{Im} Y_k(j), \end{aligned} \right\} j=o(1)n-1, k=o(1)p,$$

und

$$q := j + 2^{p-1-k}, r := j - 2^{p-1-k},$$

sowie (42) geht (47) über in

$$\left. \begin{aligned} G_{k+1}(j) &= G_k(j) + G_k(q), \\ H_{k+1}(j) &= H_k(j) + H_k(q), \end{aligned} \right\} \text{ falls } j_k = 0, \quad (47a)$$

$$\left. \begin{aligned} G_{k+1}(j) &= (G_k(r) - G_k(j)) \alpha_k(j) + (H_k(r) - H_k(j)) \alpha_k(j), \\ H_{k+1}(j) &= (H_k(r) - H_k(j)) \alpha_k(j) + (G_k(r) - G_k(j)) \alpha_k(j), \end{aligned} \right\} \text{ falls } j_k = 1. \quad (48b)$$

Nur nehmen nun an, daß wieder eine Maschinenzifferarithmetik wie in Unterausschnitt 5.4.4 eingeführt gegeben sei; Insbesondere gelten dann (38) und (43). Mit den Setzungen

$$\left. \begin{aligned} \tilde{G}_0(j) &:= \uparrow G_0(j), \\ \tilde{H}_0(j) &:= \uparrow H_0(j), \end{aligned} \right\} j=o(1)n-1, \quad (49a)$$

gehen die Transformationen (48) über in

$$\left. \begin{aligned} \tilde{G}_{k+1}(j) &:= \tilde{G}_k(j) + \tilde{G}_k(q), \\ \tilde{H}_{k+1}(j) &:= \tilde{H}_k(j) + \tilde{H}_k(q), \end{aligned} \right\} \text{ falls } j_k = 0, \quad (49b)$$

und

$$\left. \begin{aligned} \tilde{G}_{k+1}(j) &:= (\tilde{G}_k(r) - \tilde{G}_k(j)) * \tilde{U}_k(j) \approx (\tilde{H}_k(r) - \tilde{H}_k(j)) * \tilde{V}_k(j), \\ \tilde{H}_{k+1}(j) &:= (\tilde{H}_k(r) - \tilde{H}_k(j)) * \tilde{U}_k(j) \mp (\tilde{G}_k(r) - \tilde{G}_k(j)) * \tilde{V}_k(j), \end{aligned} \right\} \quad (49c)$$

falls  $j_k = 1$ .

Wir setzen noch

$$\tilde{Y}_k(j) := \tilde{G}_k(j) \mp i \tilde{H}_k(j), \quad j = o(1)n-1, \quad k = o(1)p.$$

Mit Induktion folgt dann aus (48), (49) mit Lemma 3.6

$$Y_k \subseteq \tilde{Y}_k, \quad k = o(1)p. \quad (50)$$

Aus (49) ergibt sich mit (39) und (40)

$$\left. \begin{aligned} \tilde{G}_0(j) &\subseteq G_0(j)(1 + [\epsilon]), \\ \tilde{H}_0(j) &\subseteq H_0(j)(1 + [\epsilon]), \end{aligned} \right\} \quad (51a)$$

$$\left. \begin{aligned} \tilde{G}_{k+1}(j) &\subseteq (\tilde{G}_k(j) + \tilde{G}_k(q))(1 + [\epsilon]), \\ \tilde{H}_{k+1}(j) &\subseteq (\tilde{H}_k(j) + \tilde{H}_k(q))(1 + [\epsilon]), \end{aligned} \right\} \text{ falls } j_k = 0, \quad (51b)$$

und unter Verwendung von (43)

$$\left. \begin{aligned} \tilde{G}_{k+1}(j) &\subseteq \left\{ (\tilde{G}_k(r) - \tilde{G}_k(j))u_k(1 + \delta_n[\epsilon])(1 + [\epsilon])^2 \right. \\ &\quad \left. - (\tilde{H}_k(r) - \tilde{H}_k(j))v_k(1 + \delta_n[\epsilon])(1 + [\epsilon])^2 \right\} (1 + [\epsilon]), \\ \tilde{H}_{k+1}(j) &\subseteq \left\{ (\tilde{H}_k(r) - \tilde{H}_k(j))u_k(1 + \delta_n[\epsilon])(1 + [\epsilon])^2 \right. \\ &\quad \left. + (\tilde{G}_k(r) - \tilde{G}_k(j))v_k(1 + \delta_n[\epsilon])(1 + [\epsilon])^2 \right\} (1 + [\epsilon]), \end{aligned} \right\} \quad (51c)$$

falls  $j_k = 1$ .

Im folgenden soll zunächst eine obere Schranke für die Überschätzung

$$\left. \begin{aligned} \sigma_k(j) &:= \text{span } \tilde{Y}_k(j) - \text{span } Y_k(j) \end{aligned} \right\} \quad (52)$$

$j = o(1)n-1, \quad k = o(1)p,$

1.1. Für die Überschätzung der Spanne der exakten Transformierten durch die Spanne der mit der Maschineninter-vallararithmetik berechneten hergeleitet werden.

Im Fall  $j_k=0$  ergibt sich aus (48a) und (51b)

$$\sigma_{k+1}(j) \leq s(\tilde{Y}_k(j)) + s(\tilde{Y}_k(q)) - s(Y_k(j)) - s(Y_k(q)) \\ + s((\tilde{G}_k(j)+\tilde{G}_k(q))[\epsilon]) + s((\tilde{H}_k(j)+\tilde{H}_k(q))[\epsilon]) .$$

Auf die beiden letzten Summanden der rechten Seite kann (1.2d) angewendet werden, und man erhält

$$\sigma_{k+1}(j) \leq \sigma_k(j) + \sigma_k(q) + 2\epsilon(|\tilde{G}_k(j)+\tilde{G}_k(q)| + |\tilde{H}_k(j)+\tilde{H}_k(q)|)$$

und unter Anwendung von (50) und (1.2b) folgt

$$\sigma_{k+1}(j) \leq \sigma_k(j) + \sigma_k(q) + 2\epsilon(|G_k(j)+G_k(q)| + |H_k(j)+H_k(q)| \\ + \sigma_k(j) + \sigma_k(q))$$

$$\sigma_{k+1}(j) \leq (\sigma_k(j) + \sigma_k(q))(1 + 2\epsilon) + 2\epsilon c_n^1 \quad (53)$$

$$c_n^1 := \max(|G_k(j)+G_k(j+2^{p-1-k})| + |H_k(j)+H_k(j+2^{p-1-k})|) \quad (54)$$

für alle  $j=0(1)n-1$  mit  $j_k=0$ ,  $k=0(1)p-1$ .

Mit ähnlichen Mitteln wie im Fall  $j_k=0$  erhält man im Fall  $j_k=1$

$$\sigma_{k+1}(j) \leq (\sigma_k(j) + \sigma_k(j))(|u_k(j)| + |v_k(j)|) + 2\epsilon(|G_k(j)+G_k(j)+H_k(j)+H_k(j)| + |u_k(j)| + |v_k(j)|)$$

wobei  $\tau(\epsilon) := (3+\delta_n)\epsilon + 3(1+\delta_n)(\epsilon^2 + \epsilon^3) + \delta_n \epsilon^4$  gesetzt ist.

Mit (35) ergibt sich

$$\sigma_{k+1}(j) \leq \sqrt{2}((\sigma_k(r) + \sigma_k(j))(1 + 2\tau(\epsilon)) + 2\tau(\epsilon)c_n'') \quad (55)$$

$$\text{mit } c_n'' := \max\{|G_k(j-2^{p-1-k})-G_k(j)|+|H_k(j-2^{p-1-k})-H_k(j)|\} \quad (56)$$

für alle  $j=o(1)n-1$  mit  $j_k=1, k=o(1)p-1$ .

Mit  $\sigma_k := \max_{j=o(1)n-1} \sigma_k(j), k=o(1)p-1$ , folgt hieraus und aus (53), (54)

$$\sigma_{k+1} \leq \sqrt{2} \, 2\sigma_k(1 + 2\tau(\epsilon)) + \tau(\epsilon)c_n \quad (57)$$

$$\text{mit } c_n := \sqrt{2} \, 2 \max\{c_n', c_n''\} . \quad (58)$$

Mit der Setzung (46) läßt sich jeder der Summanden der Summe in (29) betraglich durch  $n^{-1/2}\eta$  und damit die Summe betraglich durch  $n^{1/2}\eta$  abschätzen. Wegen

$$\max\{c_n', c_n''\} \leq 2\max\{|Y_k(j)| \mid j=o(1)n-1, k=o(1)p-1\}$$

folgt

$$c_n \leq 4 \sqrt{2n} \, \eta . \quad (59)$$

Wir setzen

$$\hat{\delta}_n := 3 + \delta_n . \quad (60)$$

Nach (57) wird  $\sigma_k$  durch ein Polynom vom Maximalgrad  $3k$  in  $\epsilon$  majorisiert, und wir erhalten

$$\sigma_k \leq \sum_{\mu=0}^{k-1} (2\sqrt{2})^\mu \hat{\delta}_n c_n \epsilon + (2\sqrt{2})^k (1+2k \hat{\delta}_n \epsilon) \sigma_0 + O(\epsilon^2).$$

Damit ist für  $k=o(1)p$

$$\sigma_k \leq \frac{n^{3/2}-1}{2\sqrt{2}-1} \hat{\delta}_n c_n \epsilon + n^{3/2}(1+2p \hat{\delta}_n \epsilon) \sigma_0 + O(\epsilon^2).$$

Mit (39) ergibt sich für  $j=0(1)n-1$

$$\sigma_0(j) = s(\tilde{Y}_0(j)) - s(Y_0(j))$$

$$\leq 2\epsilon |Y_0(j)| = \frac{2\epsilon}{n} |Y(j)| \leq \frac{4}{n} \epsilon \cdot n.$$

Wir fassen die Ergebnisse in dem folgenden Satz zusammen:

Satz 3.5: Für die vorgegebene Maschinenintervallarithmetik gelte (37), (38) und (43) mit der Setzung (60). Die FFT werde mit der Rechteckarithmetik nach (49) durchgeführt und liefere den Vektor  $\tilde{Y}_p = (\tilde{Y}_p(j))_{j=0}^{n-1}$ . Dann gilt für die Überschätzung

$$\sigma := \max_{j=0(1)n-1} \{s(\tilde{Y}_p(j)) - s(Y_p(j))\}$$

des exakten Vektors  $Y_p$  durch den mit der Maschinenintervallarithmetik berechneten die Schranke

$$\sigma \leq \left( \frac{n^{3/2}-1}{2^{1/2}-1} \right) c_n \delta_n + 4n^{1/2} \eta \epsilon + O(\epsilon^2), \quad (61)$$

wobei  $\eta$  nach (46) und  $c_n$  nach (58) gegeben sind. Die Konstante  $c_n$  läßt sich nach (59) abschätzen.

Eine intervallarithmetische Formulierung der Transformation (8) wurde in (24) und (25) bzw. (26) gegeben.

Wir wollen hier darauf verzichten, diese in einer Maschinenintervallarithmetik zu formulieren. Für die Akkumulation der Rundungsfehler erhält man mit analogen Abschätzungen das im folgenden Satz formulierte Ergebnis. Wir verzichten hier auf die Angabe der längeren Rechnung.

Satz 3.6: Es gelte wieder (37), (38) und (43). Die FFT werde mit der Rechteckarithmetik nach (49) und die Transformationen (24) und (25) bzw. (26) entsprechend mit der vorgegebenen Maschinenintervallarithmetik durchgeführt.

Die Konstante  $c_n$  werde nach (59) abgeschätzt.

Dann gilt für die Überschätzung der exakten Einschließungsintervalle  $[\alpha_j]$ ,  $[\beta_j]$  durch die mit der Maschinenintervallarithmetik ermittelten  $[\tilde{\alpha}_j]$ ,  $[\tilde{\beta}_j]$

$$\begin{aligned} & \max_j \{s([\tilde{\alpha}_j]) - s([\alpha_j]), s([\tilde{\beta}_j]) - s([\beta_j])\} \\ & \leq \max_j \{s([\tilde{\alpha}_j]) - s([\alpha_j]) + s([\tilde{\beta}_j]) - s([\beta_j])\} \\ & \leq (1 + \sqrt{2} + \delta_n' \epsilon) \sigma + \delta_n' \epsilon \sqrt{n} n + O(\epsilon^2), \end{aligned} \quad (62)$$

wobei  $\delta_n' := 6 + 2\sqrt{2} (5 + \delta_n)$  ist und  $\sigma$  nach (61) abgeschätzt werden kann.

Bemerkung: Sind insbesondere die Eingangsdaten  $\{\gamma_j\}$  sämtlich reelle Zahlen, so liefert die angegebene Schranke eine obere Schranke für die Akkumulation der Rundungsfehler bei der Berechnung der Einschließungsintervalle  $[\alpha_j]$ ,  $[\beta_j]$ .

#### 3.4.4 Schnelle Fouriertransformation mit der Kreisscheibenarithmetik

Wir wollen nun für die Durchführung der FFT mit der Kreisscheibenarithmetik eine im Vergleich zu Satz 3.5 günstigere Schranke herleiten.

Es sei nun  $\mathcal{Y} = (Y(j)) \in M_{n,1}(K(\mathbb{C}))$  und es gelten die Bezeichnungen (28). Nach Satz 3.2 ist dann  $\mathcal{Y}_p = n \mathcal{Y}_0$ .

Es sei

$$\begin{aligned} Y(j) &= \langle y(j), \rho(j) \rangle, \\ y(j) &= g(j) + i h(j), \\ n &= (y(j))_{j=0}^{n-1} \\ \rho &:= \max_{j=0(1)n-1} \rho(j). \end{aligned} \quad \left. \begin{aligned} & \\ & \\ & \end{aligned} \right\} j=0(1)n-1, \quad (63)$$

$$\begin{aligned} Y_k(j) &= \langle y_k(j), \rho_k(j) \rangle, \\ y_k(j) &= g_k(j) + i h_k(j), \\ n_k &= (y_k(j))_{j=0}^{n-1}, \quad k=0(1)p. \end{aligned} \quad \left. \begin{aligned} & \\ & \end{aligned} \right\} j=0(1)n-1, k=0(1)p.$$

Dann ist nach (18)

$$y_{k+1}(j) = \begin{cases} y_k(j) + y_k(q) & , \text{ falls } j_k=0, \\ (y_k(r) - y_k(j))w_k(j), & \text{ falls } j_k=1, \end{cases} \quad (64)$$

und

$$\rho_{k+1}(j) = \rho_k(j) + \rho_k(t) \quad (65)$$

$$\text{mit } t := \begin{cases} q & , \text{ falls } j_k = 0, \\ r & , \text{ falls } j_k = 1. \end{cases}$$

Bei praktischer Rechnung mit der vorgerebenen Maschinenvintervallarithmetik erhält man die Kreisscheiben

$$\tilde{Y}_k(j) = \tilde{y}_k(j), \quad \tilde{\rho}_k(j)$$

mit

$$\tilde{y}_k(j) = \tilde{g}_k(j) + i \tilde{h}_k(j).$$

Um die Einschließung  $Y_k(j) \subseteq \tilde{Y}_k(j)$  zu erhalten, setzen wir für  $j=0(1)n-1$  ( $\tilde{s}$  bezeichne die nach oben gerundete Spanne):<sup>1)</sup>

$$\tilde{g}_n(j) \in \tilde{G}_n(j), \tilde{h}_n(j) \in \tilde{H}_n(j) \text{ beliebig gewählt } ^2), \quad (66a)$$

$$\tilde{\rho}_n(j) := \sup(\tilde{s}(\tilde{G}_n(j)) \tilde{\tau} \tilde{s}(\tilde{H}_n(j)) \tilde{\tau} \uparrow \rho_n(j)), \quad (66b)$$

$$\text{wobei} \quad \left. \begin{array}{l} \tilde{G}_n(j) := \uparrow g_n(j), \\ \tilde{H}_n(j) := \uparrow h_n(j) \end{array} \right\} \quad (66c)$$

ist, und für  $k=0(1)p-1$

$$\tilde{g}_{k+1}(j) \in \tilde{G}_{k+1}(j), \tilde{h}_{k+1}(j) \in \tilde{H}_{k+1}(j) \text{ beliebig gewählt, } ^3) \quad (66d)$$

$$\tilde{\rho}_{k+1}(j) := \sup(\tilde{s}(\tilde{G}_{k+1}(j)) \tilde{\tau} \tilde{s}(\tilde{H}_{k+1}(j)) \tilde{\tau} \tilde{\rho}_k(j) \tilde{\tau} \tilde{\rho}_k(t)), \quad (66e)$$

wobei

$$\tilde{G}_{k+1}(j) := \left\{ \begin{array}{ll} \tilde{g}_k(j) \tilde{\tau} \tilde{g}_k(q) & , \text{ falls } j_k = 0, \\ (\tilde{g}_k(r) \tilde{\tau} \tilde{g}_k(j)) \tilde{*} \tilde{u}_k(j) \tilde{\tau} (\tilde{h}_k(r) \tilde{\tau} \tilde{h}_k(j)) \tilde{*} \tilde{v}_k(j), & \\ & \text{ falls } j_k = 1, \end{array} \right\} \quad (66f)$$

$$\tilde{H}_{k+1}(j) := \left\{ \begin{array}{ll} \tilde{h}_k(j) \tilde{\tau} \tilde{h}_k(q), & \text{ falls } j_k = 0, \\ (\tilde{h}_k(r) \tilde{\tau} \tilde{h}_k(j)) \tilde{*} \tilde{u}_k(j) \tilde{\tau} (\tilde{g}_k(r) \tilde{\tau} \tilde{g}_k(j)) \tilde{*} \tilde{v}_k(j), & \\ & \text{ falls } j_k = 1. \end{array} \right\} \quad (66g)$$

1) Wir unterdrücken hier die gesonderte Kennzeichnung einpunktiger Maschinenintervalle.

2) In Triplex-ALGOL beispielsweise als main-Wert von  $\tilde{G}_n(j)$  bzw.  $\tilde{H}_n(j)$ .

3) In Triplex-ALGOL beispielsweise als main-Wert von  $\tilde{G}_{k+1}(j)$  bzw.  $\tilde{H}_{k+1}(j)$ .

Dann gilt:

Satz 3.2: Es gelte wieder (37), (38) und (43) mit der Notation (60). Die FFT werde mit der Freischiebenarithmetik nach (66) durchgeführt und liefere den Vektor  $\tilde{Y}_P = (\tilde{Y}_P(j))$ . Dann gilt für die Überschätzung

$$\tau := \max_{j=0(1)n-1} \{s(\tilde{Y}_P(j)) - s(Y_P(j))\}$$

des exakten Vektors  $Y_P$  durch den mit der Maschinenintervallarithmetik berechneten

$$\tau \leq (8\hat{\delta}_n \|y\|_\infty + 3\rho)2n\epsilon + (\sqrt{2}\|y\|_\infty + \rho)4\epsilon + O(\epsilon^2).$$

Beweis: Wir wollen zunächst  $s(\tilde{G}_{k+1}(j)) - s(\tilde{H}_{k+1}(j))$  aus (66) abschätzen. Für ein Maschinenintervall  $[\tilde{a}] = [\underline{\tilde{a}}, \bar{\tilde{a}}]$  ist nach (39)

$$\tilde{s}([\tilde{a}]) \leq (\bar{\tilde{a}} - \underline{\tilde{a}})(1+\epsilon) = s([\tilde{a}]) (1+\epsilon).$$

Damit ist für  $k = 1(1)n$

$$s(\tilde{G}_k(j)) - s(\tilde{H}_k(j)) \leq (s(\tilde{G}_k(j)) + s(\tilde{H}_k(j)))(1+2\epsilon) + O(\epsilon^2). \quad (67)$$

Im Fall  $\underline{j}_k = 0$  erhält man mit den schon häufig verwendeten Abschätzungen

$$s(\tilde{G}_{k+1}(j)) + s(\tilde{H}_{k+1}(j)) \leq 2\epsilon(|\tilde{G}_k(j)| + |\tilde{H}_k(j)|) + O(\epsilon^2), \quad (68)$$

und im Fall  $\underline{j}_k = 1$

$$\begin{aligned} \tilde{G}_{k+1}(j) &= (\tilde{g}_k(r) - \tilde{g}_k(j)) \tilde{h}_k(r) (1 + \epsilon) + (1 + \epsilon) \tilde{h}_k(r) \\ &\quad - (\tilde{h}_k(r) - \tilde{h}_k(j)) \tilde{g}_k(r) (1 + \epsilon) + (1 + \epsilon) \tilde{g}_k(r) \end{aligned}$$

eine analoge Formel ist für  $\tilde{H}_{k+1}(j)$  gültig.

Damit ist mit (43) und (60)

$$\begin{aligned} & s(\tilde{G}_{k+1}(j)) + s(\tilde{H}_{k+1}(j)) \\ & \leq (|\tilde{g}_k(r) - \tilde{g}_k(j)| + |\tilde{h}_k(r) - \tilde{h}_k(j)|) 2\sqrt{2} \hat{\delta}_n \epsilon + O(\epsilon^2) \\ & \leq (|\tilde{g}_k(r) - g_k(r)| + |\tilde{g}_k(j) - g_k(j)| + |\tilde{h}_k(r) - h_k(r)| + \\ & \quad + |\tilde{h}_k(j) - h_k(j)| \\ & \quad + 2 \max_j \{ |g_k(j)| + |h_k(j)| \}) 2\sqrt{2} \hat{\delta}_n \epsilon + O(\epsilon^2). \end{aligned} \quad (69)$$

Diese Abschätzung ist nach (68) mit  $q$  anstelle von  $r$  auch im Fall  $j_k = 0$  gültig.

Nun gilt

$$|g_k(j)| + |h_k(j)| \leq \sqrt{2} \|y_k(j)\|, \quad j=0(1)n-1,$$

und somit mit der Maximumsnorm als Vektornorm

$$\max_j \{ |g_k(j)| + |h_k(j)| \} \leq \sqrt{2} \|y_k\|. \quad (70)$$

Nach (64) gilt

$$y_{k+1} = \frac{1}{n} y_k \cdot \dots \cdot y_0 y,$$

woraus mit (18) und der Zeilensummennorm als  $M$

folgt

$$\begin{aligned} \|y_{k+1}\| & \leq \frac{1}{n} \|y_k\| \cdot \dots \cdot \|y_0\| \|y\| \\ & \leq \frac{1 + \epsilon}{n} \|y_k\| \cdot \dots \cdot \|y_0\| \|y\|. \end{aligned}$$

Mittels Induktion zeigt man leicht, daß für  $k = 1(1)n$  gilt

$$\begin{aligned} & \max_j (|\tilde{p}_k(j) - p_k(j)| + |\tilde{h}_k(j) - h_k(j)|) \\ & \leq n^{3/2} \max_j (|\tilde{g}_n(j) - g_n(j)| + |\tilde{h}_n(j) - h_n(j)|) \epsilon. \end{aligned}$$

Damit folgt aus (69) mit (70), (71) und (67)

$$\tilde{s}(\tilde{G}_{k+1}(j)) \neq \tilde{s}(\tilde{H}_{k+1}(j)) \leq 8\hat{\delta}_n \|v\| \epsilon + O(\epsilon^2).$$

Aus (66e) erhalten wir

$$\tilde{p}_{k+1}(j) \leq 8\hat{\delta}_n \|v\| \epsilon + \tilde{p}_k(j)(1+2\epsilon) + \tilde{p}_k(t)(1+\epsilon) + O(\epsilon^2)$$

und mit

$$\tau_k(j) := 2(\tilde{p}_k(j) - p_k(j)), \quad j = o(1)n-1, \quad k = o(1)p,$$

wegen (65)

$$\begin{aligned} \tau_{k+1}(j) & \leq 16\hat{\delta}_n \|v\| \epsilon + \tau_k(j)(1+2\epsilon) + \tau_k(t)(1+\epsilon) \\ & \quad + 2\epsilon(2p_k(j) + p_k(t)) + O(\epsilon^2). \end{aligned}$$

Es sei  $\rho_k := \max_{j=o(1)n-1} \rho_k(j)$ , dann gilt nach (65) mit (63)

$$\rho_k \leq 2^k \rho_n \leq \rho$$

und mit  $\tau_k := \max_{j=o(1)n-1} \tau_k(j)$  folgt

$$\tau_{k+1} \leq (2+3\epsilon)\tau_k + 2(8\hat{\delta}_n \|v\| + 3\rho)\epsilon + O(\epsilon^2),$$

woraus sich mit Induktion für  $k = 1(1)p$

$$\tau_k \leq (8\hat{\delta}_n \|v\| + 3\rho) 2n\epsilon + (1+p\frac{3}{2}\epsilon)n\tau_0 + O(\epsilon^2)$$

ergibt. Aus (66c) folgt unter Verwendung von (67)

$$\begin{aligned} \tilde{S}(\tilde{G}_n(j)) - \tilde{S}(\tilde{H}_n(j)) &\leq 2\epsilon(|g_n(j)| + |h_n(j)|) + O(\epsilon^2) \\ &\leq 2\sqrt{2} \epsilon |y_n(j)| + O(\epsilon^2) \\ &= \frac{2\sqrt{2}}{n} \epsilon |y(j)| + O(\epsilon^2). \end{aligned}$$

Aus (66b) ergibt sich dann

$$\begin{aligned} \tau_o(j) &= 2(\tilde{\rho}_o(j) - \rho_o(j)) \\ &\leq \frac{2}{n} (2\sqrt{2} |y(j)| + 2\rho(j))\epsilon + O(\epsilon^2) \end{aligned}$$

und somit

$$\tau_o \leq \frac{2}{n} (2\sqrt{2} \|y\| + 2\rho)\epsilon + O(\epsilon^2).$$

Damit folgt die Behauptung. ▲

Wir übergehen hier die Angabe einer oberen Schranke für die Rundungsfehler, die bei der Ermittlung der Intervalle  $[\alpha_j]$ ,  $[\beta_j]$  entstehen, falls die FFT mit der Kreisscheibenarithmetik durchgeführt wird. Zwar lassen sich für die Akkumulation der Rundungsfehler für die Kreisscheibeneinschließung nach (21) und die Rechteckeneinschließung nach (22) leicht Schranken angeben, jedoch besteht eine Schranke für die Rundungsfehler des gesamten in Abschnitt 3.3 unter (a), (b) und (c) beschriebenen Vorgehens aus einer Vielzahl von Termen und wird sehr unübersichtlich.

Für  $\delta_n = 0$  und  $\sigma_n = 0$  bzw.  $\tau_o = 0$  enthalten die Sätze 3.5, 3.6 und 3.7 als Spezialfälle die vom Verfasser hergeleiteten Abschätzungen in [23].

### 3.4.5 Abschätzung der Konstanten $\delta_n$

Wir wollen nun die Konstanten  $\delta_n$  aus (41) abschätzen für den Fall, daß

$$\cos \xi_u, \sin \xi_u, u = o(1)2n-1, \quad (72)$$

rekursiv berechnet werden.<sup>1)</sup>

Es genügt, sich auf die Bestimmung von  $\cos \xi_u, \sin \xi_u$  für  $u = o(1)\frac{n}{n} - 1$  zu beschränken, da sich im Fall  $n = 2^p$  die restlichen aus diesen durch Vertauschen von Real- und Imaginärteil und/oder Negativsetzen des Realteils ergeben.

Es sei

$$u := \cos \pi/n, v := \sin \pi/n,$$

und

$$\left. \begin{array}{l} \tilde{U}_1 := \tilde{U} := \widetilde{\cos \pi/n} \\ \text{bzw. } \tilde{V}_1 := \tilde{V} := \widetilde{\sin \pi/n} \end{array} \right\} \quad (73)$$

bezeichne die Maschinenintervalleinschließung für den Wert des Cosinus bzw. Sinus an der Stelle  $\pi/n$ . Dann gibt es ein  $\gamma > 0$  mit

$$\tilde{U} \subseteq u(1+\gamma[\epsilon]), \tilde{V} \subseteq v(1+\gamma[\epsilon]).$$

Man erhält Einschließungen  $\tilde{U}_u, \tilde{V}_u$  für die Werte (72), indem man setzt

1) Für  $n \leq 2$  kann  $\delta_n = 0$  und für  $n = 4$  kann  $\delta_n = 1$  gewählt werden. Es sei hier deshalb  $n \geq 8$  vorausgesetzt.

$$\left. \begin{aligned} \tilde{U}_{u+1} &:= \tilde{U} * \tilde{U}_u \approx \tilde{V} * \tilde{V}_u, \\ \tilde{V}_{u+1} &:= \tilde{V} * \tilde{U}_u \mp \tilde{U} * \tilde{V}_u, \end{aligned} \right\} u = 1(1)\frac{n}{4} - 1. \quad (74)$$

Ist

$$\delta_n^* \geq \frac{1}{\epsilon} \max \left\{ \frac{s(\tilde{U}_u)}{|\cos \xi_u|}, \frac{s(\tilde{V}_u)}{|\sin \xi_u|} \right\} \quad u = 1(1)\frac{n}{4} - 1 \quad (75)$$

dann erfüllt  $\delta_n^*$  (41).

Mit den in diesem Abschnitt schon häufig verwendeten Abschätzungen erhält man für  $u = 1(1)\frac{n}{4} - 1$

$$s(\tilde{U}_u) + s(\tilde{V}_u) \leq 2^{n/8} 6.9 (2+\gamma)\epsilon + O(\epsilon^2).$$

Damit ist aber  $\delta_n^*$  nach (75) für große  $n$  unbrauchbar.

Es ist daher sinnvoll, auf eine globale Schranke, wie sie (75) darstellt, zu verzichten, und stattdessen  $\delta_n$  maschinenabhängig aus den mit der Maschinenintervallarithmetik berechneten  $\tilde{U}_u$ ,  $\tilde{V}_u$  zu ermitteln. Das erreicht man, indem man setzt

$$\delta_n := \frac{1}{\epsilon} \max \left\{ \left| \frac{\tilde{U}_u}{\cos \xi_u} - 1 \right|, \left| \frac{\tilde{V}_u}{\sin \xi_u} - 1 \right| \mid u = 1(1)\frac{n}{4} - 1 \right\}; \quad (76)$$

$\delta_n$  erfüllt dann ebenfalls (41).

Bei den im folgenden Abschnitt angegebenen Beispielen

(und in dem im Anhang aufgeführten Programm) wurden die

Einschließungen  $\tilde{U}_u$ ,  $\tilde{V}_u$  nicht nach (74), sondern mit einer intervallararithmetischen Variante des Algorithmus aus [47],

S. 437, berechnet. Dadurch konnte für größere  $n$  eine

leichte Verbesserung der Genauigkeit gegenüber (74) erreicht

werden. Für diese  $\tilde{U}_u$ ,  $\tilde{V}_u$  ist in der Tab.1  $\delta_n$  für  $2^p$ ,

$p = 3(1)10$ , aufgeführt. Man entnimmt der Tabelle, daß sich die  $\delta_n$  mit jeder Potenz von 2 etwa verdoppeln.

| $n$  | $\delta_n$ |
|------|------------|
| 8    | 0.597      |
| 16   | 0.895      |
| 32   | 1.193      |
| 64   | 1.640      |
| 128  | 2.832      |
| 256  | 4.769      |
| 512  | 9.537      |
| 1024 | 17.137     |

Tab.1 Die Konstanten  $\delta_n$  (auf drei Stellen hinter dem Dezimalpunkt aufgerundet)

### 3.4.6 Numerische Ergebnisse

In diesem Unterabschnitt geben wir für den Fall  $n = 2^p$  für eine Reihe von Beispielen die Spanne der Einschließungsintervalle für die Koeffizienten des trigonometrischen Interpolationspolynoms an. Für jedes der Beispiele wurden diese Intervalle sowohl direkt nach (2') als auch unter Einsatz der FFT mit der Rechteck- und mit der Kreisscheibenarithmetik berechnet. Das Programm für den Einsatz der Kreisscheibenarithmetik ist im Anhang aufgeführt. Für die Rechteckarithmetik wurde auf der Grundlage der ALGOL-Programme in [49] und [47], S. 437, ein Triplex-ALGOL Programm angefertigt.

Für  $n = 2^p$ ,  $p = 3(1)10$ , wurden je 20 Beispiele gerechnet.

Die Intervalle  $[\eta_k]$  wurden dabei folgendermaßen ermittelt: Es wurden je  $2n$  Zufallszahlen  $\eta_k$  aus dem Intervall  $[-1,1]$  erzeugt und

$$[\eta_k] := \eta_k + 10^{-5}[-1,1]$$

gesetzt.

Es kann also  $\alpha$  aus (31) als  $2 \cdot 10^{-5}$  gewählt werden. Nach (32) läßt sich dann die Spanne der Intervalle  $[\alpha_j]$ ,  $[\beta_j]$  für  $j=0(1)n$  abschätzen durch

$$\text{span}[\alpha_j], \text{span}[\beta_j] < \left\{ \begin{array}{l} 6.829 \\ 5.829 \\ \sqrt{n} \quad 2.415 \end{array} \right\} \cdot 10^{-5} \text{ für die } \left. \begin{array}{l} \text{Kreisscheibenarithmetik} \\ \text{Rechteckarithmetik und } \left\{ \begin{array}{l} p=3 \\ p \geq 4 \end{array} \right\} \end{array} \right\} \quad (77)$$

Für die Intervalle  $[\alpha'_j]$ ,  $[\beta'_j]$  aus (2') verifiziert man leicht

$$\text{span}[\alpha'_j], \text{span}[\beta'_j] \leq 2\alpha = 4 \cdot 10^{-5}. \quad (78)$$

Es ist ferner für  $n$  aus (46)  $n \leq 1 + 10^{-5}$ . Die vorhandene Maschinengenauigkeit  $\epsilon$  aus (37) beträgt  $\epsilon = 10^{-7}$ . In der Triplex-ALGOL 60-Version, die derzeit auf der UNIVAC 1100/81 des Rechenzentrums der Universität Freiburg i.Br. zur Verfügung steht, werden Intervalle intern mittels Näherungswert und einer unteren und einer oberen Abweichung dargestellt. Diese Darstellung besitzt den Nachteil, daß für Intervalle  $[a]$  und  $[b]$  aus (39) und (40) das Intervall  $[\epsilon]$  nicht  $10^{-7}[-1,1]$ , sondern i.a. ein breiteres Intervall ist, z.B. für Intervalle der Spanne  $10^{-4}$  ungefähr  $10^{-6}[-1,1]$ . Diese Ungenauigkeit vergrößert sich noch mit wachsender

Spanne der bei den arithmetischen Operationen verknüpften Intervalle. Im folgenden werden wir die a-priori-Fehler-schranken mit  $\epsilon = 10^{-7}$  angeben, um ein Bild von ihrem mindesten Wachstum zu geben.

In Tab. 2 sind für die direkte Berechnung nach (2') bzw. für die FFT in der Rechteckarithmetik obere Schranken für die Überschätzung der Spanne der Intervalle  $[\alpha_j']$ ,  $[\beta_j']$  bzw.  $[\alpha_j]$ ,  $[\beta_j]$  durch die Spanne der auf dem Computer berechneten Einschließungen  $[\tilde{\alpha}_j']$ ,  $[\tilde{\beta}_j']$  bzw.  $[\tilde{\alpha}_j]$ ,  $[\tilde{\beta}_j]$  angegeben. Für die direkte Berechnung nach (2') wächst die Überschätzung etwa mit  $4n$ , verdoppelt sich also etwa mit jeder Potenz von 2.

| n    | direkte Rechnung nach (2') | FFT in der Rechteckarithmetik |
|------|----------------------------|-------------------------------|
| 8    | 4.439 $10^{-8}$            | 1.815 $10^{-4}$               |
| 16   | 7.759 $10^{-8}$            | 7.889 $10^{-4}$               |
| 32   | 1.428 $10^{-5}$            | 3.343 $10^{-3}$               |
| 64   | 2.726 $10^{-5}$            | 1.474 $10^{-2}$               |
| 128  | 5.334 $10^{-5}$            | 7.158 $10^{-2}$               |
| 256  | 1.054 $10^{-4}$            | 3.946 $10^{-1}$               |
| 512  | 2.097 $10^{-4}$            | 3.055 $10^0$                  |
| 1024 | 4.175 $10^{-4}$            | 19.615 $10^0$                 |

Tab. 2 Schranken für die Überschätzung der Spannen der Einschließungsintervalle nach Satz 3.4 und Satz 3.6

Für jedes n und für jedes der je 20 Beispiele wurden nach der Berechnung der Einschließungsintervalle deren maximale Spanne

$$B_{k,n} := \begin{cases} \max_j \max\{s(\lceil \tilde{\alpha}_j^{(k,n)} \rceil), s(\lceil \tilde{\beta}_j^{(k,n)} \rceil)\} & \text{im Fall der direkten Berechnung nach (2')} \\ \max_j \max\{s(\lceil \tilde{\alpha}_j^{(k,n)} \rceil), s(\lceil \tilde{\beta}_j^{(k,n)} \rceil)\} & \text{im Fall der FFT in der Rechteck- und} \\ & \text{der Kreisscheibenarithmetik} \end{cases}$$

bestimmt. In den zweiten und dritten Spalten der Tabellen 3, 4 und 5 sind  $\min_k B_{k,n}$  bzw.  $\max_k B_{k,n}$  auf je drei Stellen hinter dem Dezimalpunkt auf- bzw. abgerundet für die direkte Rechnung nach (2') und die Durchführung der FFT in der Rechteck- und der Kreisscheibenarithmetik aufgeführt. Abb. 3 veranschaulicht diese Werte. Da im Rahmen der Zeichengenauigkeit  $\max B_{k,n}$  und  $\min B_{k,n}$  übereinstimmen, ist in Abb. 3 für diese beiden Werte und damit für alle  $B_{k,n}$ ,  $k=1(1)20$ , nur ein Wert,  $B_n$  genannt, aufgeführt.

| n    | $\min B_{k,n}$     | $\max B_{k,n}$     | ob. Schranke für $\max\{s[\tilde{\alpha}_j], s[\tilde{\beta}_j]\}$ |
|------|--------------------|--------------------|--------------------------------------------------------------------|
| 8    | 4.053 $10^{-5}$    | 4.071 $10^{-5}$    | 4.444 $10^{-5}$                                                    |
| 16   | 4.106 $10^{-5}$    | 4.125 $10^{-5}$    | 4.776 $10^{-5}$                                                    |
| 32   | 4.264 $10^{-5}$    | 4.288 $10^{-5}$    | 5.428 $10^{-5}$                                                    |
| 64   | 4.464 $10^{-5}$    | 4.626 $10^{-5}$    | 6.726 $10^{-5}$                                                    |
| 128  | 5.054 $10^{-5}$    | 5.180 $10^{-5}$    | 9.334 $10^{-5}$                                                    |
| 256  | 6.037 $10^{-5}$    | 6.772 $10^{-5}$    | 1.454 $10^{-4}$                                                    |
| 512  | 1.144 $10^{-4}$ 1) | 1.150 $10^{-4}$ 1) | 2.497 $10^{-4}$                                                    |
| 1024 | 8.945 $10^{-4}$ 1) | 8.947 $10^{-4}$ 1) | 4.575 $10^{-4}$                                                    |

Tab. 3 Ergebnisse für die direkte Berechnung nach (2')

1) Im Fall  $n=512$  wurden nur 15 und im Fall  $n=1024$  nur 3 Beispiele gerechnet, vgl. Fußnote zu Tab. 6.

| $n$  | $\min_k \beta_{k,n}$ | $\max_k \beta_{k,n}$ | ob.Schranke für<br>$\max\{s[\tilde{\alpha}_j], s[\tilde{\beta}_j]\}$ |
|------|----------------------|----------------------|----------------------------------------------------------------------|
| 8    | 5.626 $10^{-5}$      | 5.644 $10^{-5}$      | 2.398 $10^{-4}$                                                      |
| 16   | 6.747 $10^{-5}$      | 6.760 $10^{-5}$      | 3.857 $10^{-4}$                                                      |
| 32   | 7.879 $10^{-5}$      | 7.904 $10^{-5}$      | 3.580 $10^{-3}$                                                      |
| 64   | 9.265 $10^{-5}$      | 9.287 $10^{-5}$      | 1.494 $10^{-3}$                                                      |
| 128  | 1.102 $10^{-4}$      | 1.105 $10^{-4}$      | 7.186 $10^{-2}$                                                      |
| 256  | 1.327 $10^{-4}$      | 1.331 $10^{-4}$      | 3.947 $10^{-1}$                                                      |
| 512  | 1.621 $10^{-4}$      | 1.629 $10^{-4}$      | 2.546 $10^0$                                                         |
| 1024 | 2.007 $10^{-4}$      | 2.017 $10^{-4}$      | 19.616 $10^0$                                                        |

Tab. 4 Ergebnisse für die Rechteckarithmetik

| $n$  | $\min_k \beta_{k,n}$ | $\max_k \beta_{k,n}$ |
|------|----------------------|----------------------|
| 8    | 6.586 $10^{-5}$      | 6.604 $10^{-5}$      |
| 16   | 6.812 $10^{-5}$      | 6.855 $10^{-5}$      |
| 32   | 6.893 $10^{-5}$      | 6.927 $10^{-5}$      |
| 64   | 6.926 $10^{-5}$      | 6.956 $10^{-5}$      |
| 128  | 6.973 $10^{-5}$      | 6.997 $10^{-5}$      |
| 256  | 7.036 $10^{-5}$      | 7.082 $10^{-5}$      |
| 512  | 7.179 $10^{-5}$      | 7.231 $10^{-5}$      |
| 1024 | 7.492 $10^{-5}$      | 7.551 $10^{-5}$      |

Tab. 5 Ergebnisse für die Kreisscheibenarithmetik

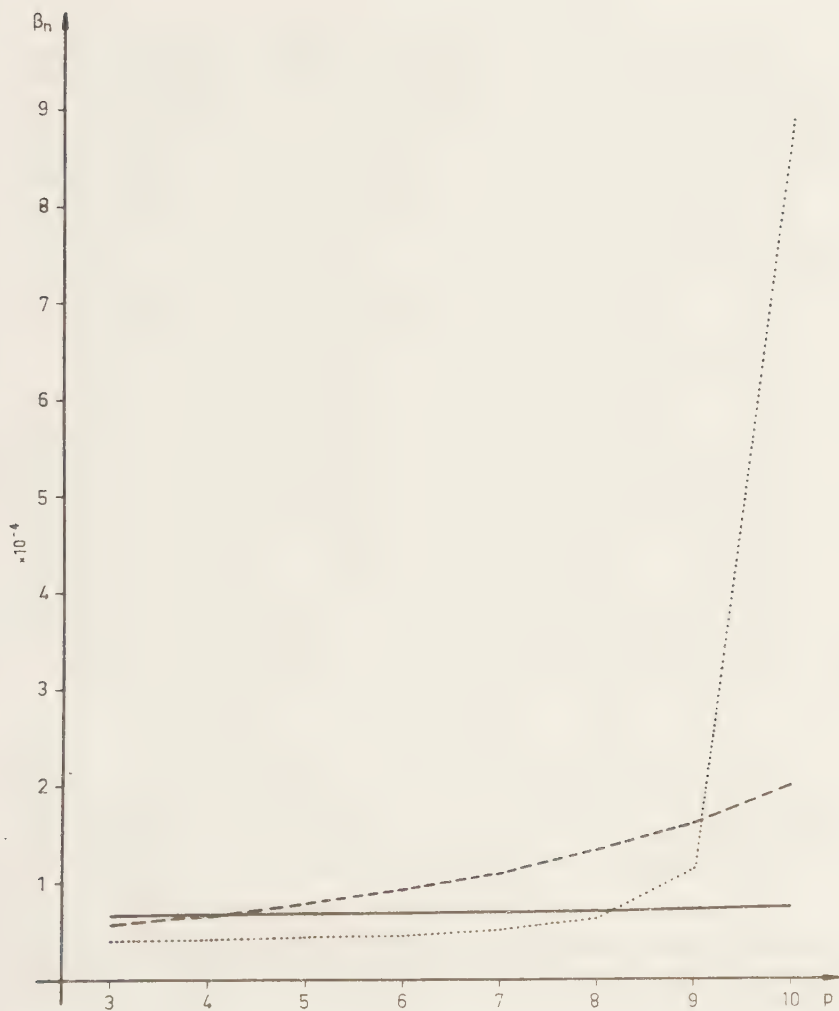


Abb. 3 Die experimentell ermittelten Spannen der Einschließungsintervalle für die Koeffizienten der trigonometrischen Interpolationspolynome für die direkte Berechnung nach (2') (punktierte Linie) und für den Einsatz der FFT in der Rechteckarithmetik (gestrichelte Linie) und in der Kreisscheibenarithmetik (durchgezogene Linie).

A-priori-Schranken für die Spannen der Einschließungsintervalle erhält man, indem man zu den Werten aus Tab. 2 die entsprechenden Werte aus (77) und (78) addiert. Die auf diese Weise ermittelten Schranken sind in der 3. Spalte von Tab. 3 bzw. Tab. 4 angegeben. Im Falle der FFT in der Rechteckarithmetik werden diese Schranken jedoch in praxi weit unterschritten. Im Falle der direkten Berechnung überschätzt die obere Schranke für  $\max\{s[\alpha_j^!], s[\beta_j^!]\}$  für  $n < 1024$  die experimentell ermittelten Werte nur wenig; für  $n = 1024$  wird sie sogar von den drei erhaltenen Werten überschritten. Dies hat seinen Grund in der schon erwähnten Überschätzung infolge der internen Darstellung von Intervallen sowie der sich damit auch für  $n \geq 1024$  bemerkbar machenden Vernachlässigung der  $\epsilon$ -Glieder höherer als 1. Ordnung (im Beweis zu Satz 3.4 schätzt man die  $\epsilon^2$ -Glieder durch  $(8/3 \eta n^2 + O(n)) \epsilon^2$  ab).

Vergleicht man die Werte der zweiten und der dritten Spalte der Tab. 4 mit den entsprechenden Werten in Tab. 5, so zeigt sich, daß im Fall  $n < 16$  die Rechteckarithmetik die schärferen Schranken liefert. Nach den Sätzen 3.1 und 3.2 tritt zwar bei beiden Arithmetiken keine Überschätzung des Produktes  $m \eta$  ein, im Fall der Kreisscheibenarithmetik bewirkt jedoch für  $n \geq 16$  die Kreisscheibeneinschließung vor der Fouriertransformation und die Rechteckeinschließung der Transformatierten eine Überschätzung der optimalen Intervalleinschließung von  $\{c\}$ , die in etwas größeren Spannen der Einschließungsintervallen ihren Niederschlag findet. Für  $n \geq 32$  liefert der Einsatz der Kreisscheibenarithmetik

für wachsendes  $n$  jedoch eine wachsende Verbesserung gegenüber der Rechteckarithmetik. Allerdings wird dieser Vorteil der Kreisscheibenarithmetik mit etwa doppelt so langer Ausführungszeit (execution time) gegenüber der Rechteckarithmetik erkaufte (vgl. Tab. 6).

Werden die Einschließungen nach (2') berechnet, dann ist für  $n \leq 256$  bzw.  $n \leq 512$  die Genauigkeit höher als beim Einsatz der FFT in der Kreisscheibenarithmetik bzw. in der Rechteckarithmetik. Für  $n > 512$  verliert die direkte Methode jedoch sehr schnell an Genauigkeit. Daß die direkte Berechnung für große  $n$  praktisch unbrauchbar ist, zeigt die erste Spalte von Tab. 6. Während sich die Ausführungszeit für die FFT in einer der beiden Arithmetiken mit jeder Potenz von 2 etwa verdoppelt, vervierfacht sich die Zeit bei der direkten Berechnung.

#### Zusammenfassung:

Für kleine  $n, n \leq 32$ , kann man mit dem Vorteil einer erhöhten Genauigkeit die direkte Berechnung anwenden. Wird für  $n \geq 32$  das Hauptgewicht auf einfache Programmierung und kurze Rechenzeit gelegt, so benutze man die FFT in der Rechteckarithmetik; steht dagegen eine möglichst "enge" Einschließung im Vordergrund, so verwende man die FFT in der Kreisscheibenarithmetik.

| n   | direkte Berechnung | FFT in der         |                         |
|-----|--------------------|--------------------|-------------------------|
|     |                    | Rechteckarithmetik | Kreisscheibenarithmetik |
| 8   | 1.533              | 0.873              | 1.422                   |
| 16  | 6.764              | 1.780              | 3.826                   |
| 32  | 26.743             | 4.041              | 7.886                   |
| 64  | 105.975            | 8.947              | 19.448                  |
| 128 | 424.289            | 19.514             | 42.698                  |
| 256 | 1695.685           | 44.083             | 94.011                  |
| 512 | 1 - 1 std. 1)      | 95.768             | 203.390                 |
|     | 30 min.            | 204.198            | 440.623                 |

Tab. 6 Ausführungszeit (execution time) in sec. zur Bestimmung von Einschließungen für die Koeffizienten der trigonometrischen Interpolationspolynome für jeweils 20 Wertevorgaben <sup>1)</sup>  
 $[r_k], k = 0(1)2n-1$

1) Im Fall der direkten Berechnung nach (2') wurde für  $n = 512$  im 16. und für  $n = 1024$  schon im 4. Beispiel abgebrochen, da die Zeitschranke von 1 std. 30 min. überschritten wurde.

## Kapitel 4 Intervallinterpolation mit Splines

### 4.1 Vorbemerkungen

Führt man bei der Interpolation einer stetigen Funktion durch Polynome <sup>1)</sup> nacheinander immer mehr Stützstellen ein und interpoliert mit Polynomen von entsprechend höherem Grad, so ist bekanntlich i.a. die punktweise und/oder gleichmäßige Konvergenz dieser Interpolationspolynome gegen die zu interpolierende Funktion nicht gesichert. Es sei hier an die Ergebnisse von Faber und Marcinkiewicz erinnert (vgl. [10], S. 11-13). Wesentlich besseres Konvergenzverhalten zeigen die sog. Splines, dies sind Funktionen, die sich zwischen den Stützstellen aus Polynomen von bestimmtem Maximalgrad zusammensetzen und die an den Stützstellen stetig differenzierbar von vorgeschriebener Ordnung sind.

Wir wollen uns im zweiten Abschnitt des vorliegenden Kapitels kurz mit den kubischen Splines befassen. Im dritten Abschnitt werden wir dann auf Probleme eingehen, die sich bei der Darstellung einer interpolierenden Funktion bzgl. eines Basis aus allgemeineren Spline-Funktionen, den sog. B-Splines, ergeben.

---

<sup>1)</sup> Fortan sollen unter Polynomen stets die algebraischen Polynome verstanden werden.

In beiden Fällen führt das Interpolationsproblem auf ein lineares Intervallgleichungssystem

$$[a]r = [c]$$

mit einer Intervallmatrix  $[a] \in M_{n,n}(II(\mathbb{R}))$  und rechter Seite  $[c] \in II(\mathbb{R}^n)$ .

Gesucht ist die Lösungsmenge  $\{r\}$  dieses Gleichungssystems

$$\{r\} := \{r \in \mathbb{R}^n \mid \exists a \in [a] : a r \in [c]\}.$$

I.a. gilt  $\{r\} \notin II(\mathbb{R})$ , und man versucht daher, einen Intervallvektor  $[b] \in II(\mathbb{R}^n)$  mit komponentenweise möglichst kleiner Spanne und  $\{r\} \subseteq [b]$  zu bestimmen.

Dies ist mit Hilfe des Intervall-Gauß-Algorithmus möglich, falls dieser durchführbar ist. Im folgenden Unterabschn. 4.1.1 werden wir uns mit der Durchführbarkeit des Intervall-Gauß-Algorithmus befassen. Die Frage nach der Durchführbarkeit wird dort jedoch nicht abschließend behandelt werden, noch nicht einmal so weit es im Rahmen dieser Arbeit erforderlich ist. Die im Unterabschn. 4.1.1 angegebenen Definitionen und Kriterien werden ausreichen, um die Abschn. 4.2 bei der Intervallinterpolation mit kubischen Splines auftretenden linearen Intervallgleichungssysteme zu behandeln. Die Fragestellung des Abschn. 4.3 führt uns auf andere Matrizen, nämlich auf die sog. totalniehtnegativen Matrizen. Da die Hilfsmittel des Unterabschn. 4.1.1 für die Behandlung der dort auftretenden Probleme nicht ausreichen, wird die Frage nach der Durchführbarkeit des Intervall-Gauß-Algorithmus in Unterabschn. 4.4.3 noch einmal aufgegriffen. Dort beschäftigen wir uns mit weiteren Klassen von Intervallmatrizen, insbesondere

mit den totalnichtnegativen und mit speziellen invers-isotonen Intervallmatrizen.

#### 4.1.1 Die Durchführbarkeit des Intervall-Gauß-Algorithmus für einige Klassen von Intervallmatrizen

Der Intervall-Gauß-Algorithmus, im folgenden mit IGA abgekürzt, stellt die unmittelbare Übertragung des reellen Gauß-Algorithmus auf Intervalle dar: Es werden alle auftretenden reellen Größen durch Intervalle und alle reellen arithmetischen Operationen durch die entsprechenden intervallarithmetischen Operationen ersetzt (vgl. [2], Abschnitt 15). Wie beim reellen Gauß-Algorithmus sind gegebenenfalls noch Zeilen- und Spaltenvertauschungen durchzuführen. Die folgenden Definitionen verstehen sich in Abhängigkeit von einer festgewählten Matrix  $[a]$ . Wir sprechen vom IGA ohne Pivotisierung, wenn in jedem Eliminationsschritt keine Zeilen- und Spaltenvertauschungen vorgenommen werden und wenn stets  $[a_{jj}^{(j-1)}]$  (i.e. das Element, das in der 1. Zeile und 1. Spalte der Restmatrix steht) als Pivotelement gewählt wird. Wir nennen den IGA ohne Pivotisierung durchführbar, falls  $[a_{jj}^{(j-1)}]$  stets kein Nullintervall ist ( $j = 1(1)n$ ).

Für einige Klassen von Intervallmatrizen ist der IGA stets ohne Pivotisierung durchführbar. Eine ziemlich umfassende Klasse möchten wir hier angeben, da sie für das folgende

wichtig ist.

Wir definieren dazu:

Definition (irreduzibel, invers-isoton, streng diagonal-  
dominant, irreduzibel diagonaldominant,  
M-Matrix, H-Matrix):

Eine Matrix  $a = (a_{ij}) \in M_{n,n}(\mathbb{R})$  heißt

irreduzibel, falls zu jeder nichtleeren Teilmenge I von

$I := \{1, 2, \dots, n\}$  mit  $I \neq N$  es ein  $a_{ij} \neq 0$  mit  $i \in I$   
 und  $j \in N \setminus I$  gibt;

invers-isoton, falls  $a$  regulär ist und  $a^{-1} \geq 0$  gilt;

streng diagonaldominant, falls gilt

$$\sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| < |a_{ii}|, \quad i=1(1)n;$$

irreduzibel diagonaldominant, falls  $a$  irreduzibel ist und  
 falls gilt

$$\sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| \leq |a_{ii}|, \quad i=1(1)n,$$

mit strenger Ungleichung für mindestens ein  $i$ ;

M-Matrix, falls  $a$  invers-isoton ist und falls gilt

$$a_{ij} \leq 0, \quad i, j = 1(1)n, \text{ mit } i \neq j;$$

H-Matrix, falls  $\lambda(a) = (z_{ij})$  mit

$$z_{ij} := \begin{cases} |a_{ij}| & \text{für } i=j \\ -|a_{ij}| & \text{für } i \neq j \end{cases}, \quad i, j = 1(1)n,$$

eine M-Matrix ist.

- Bemerkungen: 1.) Ob eine Matrix irreduzibel ist oder ob nicht, überprüft man häufig bequem mit Hilfe eines notwendigen und hinreichenden Kriteriums, welches den Begriff des Graphen einer Matrix verwendet (s. [51], S. 18-20).
- 2.) Jede M-Matrix ist insbesondere auch eine H-Matrix.
- 3.) Jede streng diagonaldominante bzw. irreduzibel diagonaldominante Matrix ist eine H-Matrix (vgl. [52]).

Definition (irreduzible/invers-isotone/streng diagonaldominante/irreduzibel diagonaldominante Intervallmatrix, H-Matrix(-Intervall), H-Matrix(-Intervall)):

Eine Intervallmatrix  $[a]$  heißt irreduzibel, invers-isoton, streng diagonaldominant, irreduzibel diagonaldominant, M-Matrix(-Intervall), H-Matrix(-Intervall), falls jede Matrix  $a \in [a]$  die entsprechende Eigenschaft besitzt.

Es gelten die folgenden Kriterien:

Lemma 4.1: Eine Intervallmatrix  $[a] = [\underline{a}, \overline{a}] = ([a_{ij}])_{i,j=1, \dots, n}^{(-)}$  ist

(a) irreduzibel genau dann, wenn  $\text{sign } [a]$  irreduzibel ist;

(b) invers-isoliert genau dann, wenn  $\underline{a}$  ,  $\bar{a}$  invers-isoliert sind;

(c) streng diagonaldominant genau dann, wenn gilt

$$\sum_{\substack{j=1 \\ j \neq i}}^n |[a_{ij}]| < \text{mig}[a_{ii}] \quad , \quad i=1(1)n;$$

(d) irreduzibel diagonaldominant genau dann, wenn  $\text{sigr}[a]$  irreduzibel ist und wenn gilt

$$\sum_{\substack{j=1 \\ j \neq i}}^n |[a_{ij}]| \leq \text{mig}[a_{ii}] \quad , \quad i=1(1)n \quad ,$$

mit strenger Ungleichung für mindestens ein  $i$  ;

(e) eine M-Matrix genau dann, wenn  $\underline{a}$  eine M-Matrix und  $\bar{a}_{ij} \leq 0$  für  $i, j= 1(1)n$  mit  $i \neq j$  ist;

(f) eine H-Matrix genau dann, wenn  $\lambda([a]) = (z_{ij})$  mit

$$z_{ij} := \begin{cases} \text{mig}[a_{ii}] & \text{für } i = j \\ -|[a_{ij}]| & \text{für } i \neq j \end{cases}$$

eine M-Matrix ist.

Bemerkungen: 1.) (a), (c) und (d) folgert man aus den Definitionen; zu (b) siehe [31], S.240, zu (e) [51], S. 85; (f) ergibt sich mittels (e).

2.) Wieder gilt, daß jede streng diagonaldominante und jede irreduzibel diagonaldominante Intervallmatrix auch eine H-Matrix ist.



föhrbar ist, wenn er auf spezielle tridiagonale, irreduzibel diagonaldominante Intervallmatrizen angewendet wird. Aus Bem. 2.) zu Lemma 4.1 folgt, daB der IGA auch für beliebige irreduzibel diagonaldominante Intervallmatrizen ohne Pivotisierung durchföhrbar ist.

#### 4.2 Kubische Splines

Wir notieren zunächst einige gängige Definitionen und einen wohlbekannten Satz (vgl. [10], Kapitel 1).

Definition (Zerlegung, Schrittweite der Zerlegung, kubischer Interpolationsspline):

Im Intervall  $[a,b]$  seien Stützstellen  $x_k, k=1(1)n$ ,  $n \geq 3$ , gegeben mit

$$a = x_1 < x_2 < \dots < x_n = b .$$

Den Vektor  $\tau = (x_1, x_2, \dots, x_n)^t$  nennen wir Zerlegung von  $[a,b]$  und

$$h_k := x_{k+1} - x_k, \quad k = 1(1)n-1 ,$$

die Schrittweite der Zerlegung  $\tau$ .

Die Funktion  $s_{\tau, \eta} : [a,b] \rightarrow \mathbb{R}$  heißt kubischer Interpolationsspline bzgl. der Zerlegung  $\tau$  zur Wertevorgabe

$\eta = (y_1, \dots, y_n)^t \in \mathbb{R}^n$ , wenn gilt

- (a)  $s_{\tau, \eta}$  stimmt auf  $[x_k, x_{k+1}]$  mit einem kubischen Polynom überein,  $k = 1(1)n-1$ ,

$$(b) \ s_{\mathbf{r},\mathbf{y}} \in C^2[a,b] \ ,$$

$$(c) \ s_{\mathbf{r},\mathbf{y}}(\mathbf{x}_k) = y_k \ , \ k = 1(1)n \ .$$

Die ersten bzw. die zweiten Ableitungen des Interpolationssplines an den Stützstelle  $\mathbf{x}_i$  nennt man Steigungen bzw. Momente. Für die Steigungen, im folgenden mit  $m_i$  bezeichnet, bzw. für die Momente läßt sich ein lineares Gleichungssystem aufstellen. Nimmt man noch geeignete Randbedingungen hinzu, z.B. durch Vorgabe der ersten und zweiten Ableitung des Splines in den Randpunkten  $a$  und  $b$ , so ist der Interpolationsspline eindeutig bestimmt und kann mithilfe der Lösung des Gleichungssystems für die Steigungen bzw. für die Momente bestimmt werden. Wir verwenden hier nur den Ansatz für die Steigungen, da dieser bei der praktischen Rechnung gewisse Vorteile gegenüber dem Ansatz für die Momente besitzt.

Satz 4.2: (vgl. [10], S. 20/21) Zu  $u_1, d_1, r_n, d_n \in \mathbb{R}$  mit

$$|u_1|, |r_n| < 2 \quad ( )$$

gibt es genau einen kubischen Interpolationsspline  $s_{\mathbf{r},\mathbf{y}}$  bzgl. der Zerlegung  $\mathbf{r}$  und der Wertevorgabe  $\mathbf{y} \in \mathbb{R}^n$ , der den Randbedingungen

$$2m_1 + u_1 m_2 = d_1$$

und

$$r_n m_{n-1} + 2m_n = d_n$$

genügt. Berechnet man  $m_i$ ,  $i = 1(1)n$ , als Lösung von

$$\begin{pmatrix} 2 & u_1 & & & \\ r_2 & 2 & u_2 & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & r_{n-1} & 2 & u_{n-1} & \\ & & r_n & 2 & \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_{n-1} \\ m_n \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_{n-1} \\ d_n \end{pmatrix} \quad (2)$$

mit

$$r_i := \frac{h_i}{h_{i-1} + h_i}, \quad u_i := 1 - r_i,$$

$$d_j := 3(u_i \frac{y_{i+1} - y_i}{h_i} + r_i \frac{y_i - y_{i-1}}{h_{i-1}}), \quad i = 2(1)n-1,$$

so erhält man  $s = s_{r,q}$  aus

$$\left. \begin{aligned} s|_{[x_i, x_{i+1}]}(x) &= y_i + m_i(x - x_i) + \left( 3 \frac{y_{i+1} - y_i}{h_i^2} - \frac{m_{i+1} + 2m_i}{h_i} \right) (x - x_i)^2 \\ &\quad + \left( -2 \frac{y_{i+1} - y_i}{h_i^3} + \frac{m_{i+1} + m_i}{h_i^2} \right) (x - x_i)^3. \end{aligned} \right\} \quad (3)$$

Bemerkungen: 1.) Es seien  $y'_1, y''_1, y'_n, y''_n$  irgendwelche Näherungen für die ersten bzw. zweiten Ableitungen des gesuchten Interpolationssplines in den Randpunkten  $a, b$ . Dann sind für die beiden in der Praxis häufig auftretenden Randbedingungen

$$\left. \begin{aligned} 2m_1 &= 2y'_1 \\ 2m_n &= 2y'_n \end{aligned} \right\} \quad \text{Vorgabe der 1. Ableitung}$$

und

$$\left. \begin{aligned} 2m_1 + m_2 &= 3 \frac{y_2 - y_1}{h_1} - \frac{h_1}{2} y''_1 \\ m_{n-1} + 2m_n &= 3 \frac{y_n - y_{n-1}}{h_{n-1}} + \frac{h_{n-1}}{2} y''_n \end{aligned} \right\} \quad \begin{array}{l} \text{Vorgabe der} \\ \text{2. Ableitung} \end{array}$$

die Voraussetzungen (1) des Satzes erfüllt.

2.) Man kann auch verlangen, daß der Interpolationsspline über das Intervall  $[a, b]$  hinaus periodisch mit der Periode  $b-a$  fortgesetzt werden soll. Aus dieser Forderung ergibt sich  $m_1 = m_n$  und  $m_2 = m_{n+1}$ . Man erhält dann ein Analogon zu Satz 4.1, wobei allerdings das (2) entsprechende Gleichungssystem nur noch "fast tridiagonal" ist.

Geht man nun zu einem Intervallinterpolationsproblem vom Typ II über, so sind die Stützstelle  $x_k$  bzw. die  $y$ -werte  $y_k$  durch Intervalle  $[x_k]$  bzw.  $[y_k]$  zu ersetzen und das reelle lineare Gleichungssystem (2) geht über in das folgende lineare Intervallgleichungssystem zur Bestimmung von Schranken  $\underline{m}_i, \bar{m}_i$  für die Steigungen  $m_i$

$$\underbrace{\begin{pmatrix} [u_1] & & & \\ [r_2] & 2 & [u_2] & 0 \\ & \ddots & \ddots & \ddots \\ 0 & [r_{n-1}] & 2 & [u_{n-1}] \\ & & [r_n] & 2 \end{pmatrix}}_{=: [a]} \cdot \begin{pmatrix} m_1 \\ \vdots \\ m_{n-1} \\ m_n \end{pmatrix} = \begin{pmatrix} [d_1] \\ \vdots \\ [d_{n-1}] \\ [d_n] \end{pmatrix} \quad (4)$$

mit

$$\left. \begin{aligned} [r_i] &:= \frac{[h_i]}{[h_i] + [h_{i-1}]}, [u_i] := \frac{[h_{i-1}]}{[h_i] + [h_{i-1}]}, \\ [d_i] &:= 3 \left( [u_i] \frac{[y_{i+1}] - [y_i]}{[h_i]} + [r_i] \frac{[y_i] - [y_{i-1}]}{[h_{i-1}]} \right), \end{aligned} \right\} i=2(\dots)n-1, \quad (5)$$

wobei gesetzt ist

$$[h_k] := [x_{k+1}] - [x_k], \quad k=1(\dots)n-1.$$

Ist ein Intervallvektor  $([m_i]) = ([\underline{m}_i, \bar{m}_i])$  bestimmt, der die Lösungsmenge des Intervallgleichungssystems (4) enthält, dann liefert die Ersetzung der reellen Größen  $x_i, y_i, m_i$  durch die Intervalle  $[x_i], [y_i], [m_i]$  in (3) einen Funktionsstreifen, der sämtliche kubischen Interpolationssplines  $s_{r,n}$  bzgl. der Zerlegung  $r \in ([x_i])$  und der Wertevorgabe  $n \in ([y_i])$  enthält.

Mit Lemma 4.1 (c) und (d) ergeben sich Bedingungen, unter denen der IGA, angewendet auf  $[a]$  aus (4) durchführbar ist. Ist beispielsweise

$$|[u_i]|, |[r_n]| < 2, \quad (6)$$

$$\bar{h}_i + \bar{h}_{i-1} < 2(\underline{h}_i + \underline{h}_{i-1}), \quad i=2(1)n-1, \quad (7)$$

dann ist  $[a]$  nach Lemma 4.1 (c) streng diagonaldominant und nach Satz 4.1 der IGA ohne Pivotisierung durchführbar. Die Bedingung (7) ist - wie man leicht durch Einsetzen der Eckpunkte von  $[x_{i-1}], [x_i], [x_{i+1}]$  in (7) verifiziert - äquivalent zu

$$\text{span}[x_{i-1}] + 3 \text{span}[x_i] + \text{span}[x_{i+1}] < \underline{x}_{i+1} - \bar{x}_{i-1},$$

d.h. die Spannen der Intervalle  $[x_i]$  sind beschränkt durch die Abstände der Intervalle voneinander.

### 4.3 Bestimmung von Schranken für die B-Spline-Koeffizienten einer interpolierenden Spline-Funktion

In Verallgemeinerung der im vorausgegangenen Abschnitt eingeführten kubischen Splines betrachtet man Funktionen, die

sich stückweise aus Polynomen (gegebenenfalls höheren als dritten Grades) zusammensetzen und die stetige Übergänge vorgeschriebener Ordnung besitzen. Eine Basis eines derartigen Funktionenraumes bilden die sog. B-Splines. Auf diese Basiseigenschaft möchten wir hier nicht eingehen und verweisen auf [11].

Wir werden vielmehr untersuchen, welche Schwierigkeiten auftreten, wenn Schranken für die Koeffizienten des Interpolationssplines bzgl. der B-Spline-Basis bestimmt werden sollen. Dazu notieren wir zunächst einige Definitionen und Eigenschaften der B-Splines in Anlehnung an die Einführung in [11].

Es sei  $p \geq 1$  und  $t = (\tau_i)_{i=1}^{n+p}$  eine monoton wachsende Folge reeller Zahlen mit

$$\tau_i < \tau_{i+p}, \quad i = 1(1)n.$$

Definition ((normalisierter) B-Spline der Ordnung  $p$   
 $\$_{p,t}$ , Spline der Ordnung  $p$ ):

Die Funktion

$$b_{i,p,t}(x) := (\tau_{i+p} - \tau_i) [\tau_i, \dots, \tau_{i+p}] (t-x)_+^{p-1},$$

$$x \in \mathbb{R} \setminus \{\tau_j \mid j = 1(1)n\},$$

$$i = 1(1)n,$$

heißt  $i$ -ter ((normalisierter) B-Spline der Ordnung  $p$  zur Knotenfolge  $t$ . Dabei bezeichnet  $[\tau_i, \dots, \tau_{i+p}] (t-x)_+^{p-1}$  die  $p$ -te dividierte Differenz an den Stellen  $\tau_i, \dots, \tau_{i+p}$  der Funktion in  $t$

$$(t-x)_+^{p-1} := \begin{cases} (t-x)^{p-1} & \text{für } t \geq x \\ 0 & \text{für } t < x \end{cases} \quad (0^0 := 1) ,$$

für festgehaltenes  $x$  .

Wie üblich setzt man  $b_{i,p,t}$  an den Stellen  $\tau_j$  rechtsstetig fort

$$b_{i,p,t}(\tau_j) := b_{i,p,t}(\tau_j^+), \quad j = 1(1)n .$$

Für  $b_{i,p,t}$  schreiben wir auch abkürzend  $b_{i,p}$  oder  $b_i$  .

Ein Element der Menge

$$S_{p,t} := \left\{ \sum_{i=1}^n \alpha_i b_{i,p,t} \mid \alpha_i \in \mathbb{P}, i = 1(1)n \right\}$$

heißt Spline oder Spline-Funktion der Ordnung  $p$  zur Knotenfolge  $t$  .

Die B-Splines haben die Eigenschaft, daß sie einen "kleinen" Träger besitzen. Genauer gilt:

Lemma 4.3 (vgl. [11], S.109-110): Für die oben definierten B-Splines ist

$$\left. \begin{aligned} (a) \quad & b_i(x) = 0 \quad \text{für } x \notin [\tau_i, \tau_{i+p}] , \\ (b) \quad & b_i(x) > 0 \quad \text{für } x \in ]\tau_i, \tau_{i+p}[ , \end{aligned} \right\} \quad i = 1(1)n ,$$

$$(c) \quad \sum_{i=1}^n b_i(x) = \sum_{i=j-p+1}^j b_i(x) = 1 \quad \text{für alle } x \in ]\tau_j, \tau_{j+1}[ ,$$

$$j = p(1)n .$$

Die B-Splines lassen sich rekursiv berechnen:

Mit

$$b_{i,1}(x) = \begin{cases} 1 & \text{für } \tau_i \leq x < \tau_{i+1} \\ 0 & \text{sonst} \end{cases} \quad (8)$$

ist

$$b_{i,p}(x) = \frac{x - \tau_i}{\tau_{i+p-1} - \tau_i} b_{i,p-1}(x) + \frac{\tau_{i+p} - x}{\tau_{i+p} - \tau_{i+1}} b_{i+1,p-1}(x). \quad (9)$$

Es seien nun wieder Stützstellen  $x_k$  und -werte  $y_k$ ,  $k=1(1)n$ , vorgegeben. Die Bestimmung einer Funktion  $f \in \mathcal{S}_{p,1}$  mit  $f(x_k) = y_k$ ,  $k=1(1)n$ , führt auf das lineare Gleichungssystem

$$\sum_{i=1}^n \alpha_i b_i(x_k) = y_k, \quad k = 1(1)n, \quad (10)$$

für die B-Spline-Koeffizienten  $\alpha_i$  der interpolierenden Spline-Funktion  $f$ . Die Koeffizientenmatrix  $b = (b_{ik}) = (b_i(x_k))$  dieses Gleichungssystems ist totalnichtnegativ, d.h. ihre sämtlichen Minoren sind nichtnegativ;  $b$  ist regulär genau dann, wenn gilt

$$\tau_k < x_k < \tau_{k+p}, \quad k = 1(1)n. \quad (11)$$

Unter dieser Bedingung ist  $b$  wegen Lemma 4.3 (a) eine Bandmatrix mit Bandbreite kleiner als  $p$ , d.h. es ist  $b_{ik} = 0$  für  $|i - k| \geq p$ . Ferner ist, falls (11) erfüllt ist, das lineare Gleichungssystem (10) mit dem Gauß-Algorithmus ohne Pivotsuche auflösbar. Für den Gauß-Algorithmus, angewandt auf eine totalnichtnegative, reguläre Matrix, also insbesondere auch auf  $b$ , lassen sich Schranken für die Rundungsfehler angeben, die geringer ausfallen, als diejenigen für beliebige Matrizen, sogar wenn bei diesen Pivotisierung angewendet wird (s. [12]).

Liegt nun ein Intervallinterpolationsproblem vom Typ II vor, dann variieren die B-Spline-Koeffizienten  $\alpha_i$  der interpolierenden Spline-Funktion  $f$  in Abhängigkeit von den Stützstellen  $x_k \in [x_k]$  und den Stützwerten  $y_k \in [y_k]$ . Es stellt sich die Frage nach einschließenden Intervallen  $[\alpha_i]$  für die Koeffizienten  $\alpha_i$ , derart daß für  $f \in \mathcal{S}_{p,1}$  mit  $f(x_k) \in [y_k]$ ,  $k = 1(1)n$ ,

$$f(x) \in \sum_{i=1}^n [\alpha_i] b_i(x)$$

gilt. Nach Lemma 4.3 (b), (c) ist dann

$$f(x) \in \left[ \sum_{i=j-p+1}^j \alpha_i b_i(x), \sum_{i=j-p+1}^j \bar{\alpha}_i b_i(x) \right]$$

$$\text{für alle } x \in ]\tau_j, \tau_{j+1}[.$$

Im Spezialfall, daß die Stützstellen sämtlich reell sind, also beim Intervallinterpolationsproblem vom Typ I, ist die optimale Einschließung (2.9) anwendbar, es treten also keine Schwierigkeiten auf.

Um für den Fall beliebiger Intervalle  $[x_k]$  ein lineares Intervallgleichungssystem für die Intervalle  $[\alpha_i]$  aufstellen zu können, benötigen wir Intervalle  $[b_{ik}]$ , welche  $b_i[x_k] = \{b_i(x) | x \in [x_k]\}$ , den Wertebereich von  $b_i$  auf  $[x_k]$ , enthalten. Derartige Intervalle sollen im folgenden bestimmt werden.

Nach (11) ist es sinnvoll,

$$\tau_k < [x_k] < \tau_{k+p}, \quad k = 1(1)n, \quad (12)$$

zu fordern. Weiter wollen wir der Einfachheit halber annehmen, daß gilt

$$\tau_i \notin [x_k] \quad i = 2(1)n+p-1, k=1(1)n.$$

Die Rekursionsformeln (8), (9) lassen sich derart verallgemeinern, daß sie Intervalle  $[b_{ik}]$  mit der gewünschten Einschließungseigenschaft liefern:

Rekursiv werden Intervalle  $[b_{ik}^{(j)}]$  für  $i, k=1(1)n$ ,  $j=1(1)p$ , bestimmt:

$$[b_{ik}^{(1)}] := \begin{cases} 1 & , \text{ falls } \tau_i < [x_k] < \tau_{i+1} \text{ gilt,} \\ 0 & \text{sonst} \end{cases} \quad (3)$$

und für  $j=2(1)p$

$$[b_{ik}^{(j)}] := \frac{[x_k] - \tau_i}{\tau_{i+j-1} - \tau_i} [b_{ik}^{(j-1)}] + \frac{\tau_{i+j} - [x_k]}{\tau_{i+j} - \tau_{i+1}} [b_{i+1,k}^{(j-1)}]^{1)} \quad (4)$$

Aufgrund der Inklusionsisotonie der intervallararithmetischen Operationen gilt dann für  $i, k = 1(1)n$  mit

$$[b_{ik}] := [b_{ik}^{(p)}]$$

die Inklusion

$$b_i([x_k]) \subseteq [b_{ik}] \quad (15)$$

Mit Induktion ergibt sich ferner, daß die Matrix  $[b] = ([b_{ik}])$  die gleiche Vorzeichenstruktur wie die Matrix  $b$  des korrespondierenden reellen Problems besitzt, es ist nämlich

$$\text{also} \quad \begin{cases} \text{entweder } [b_{ik}] = 0 \text{ oder } [b_{ik}] > 0, & i, k=1(1)n, \\ \text{sign } [b] & = \text{sign } b. \end{cases} \quad (16)$$

---

1) Es ist  $[b_{n+1,k}^{(j)}] := 0$  gesetzt,  $j = 1(1)n-1$ .

I.a. gilt in (15) die echte Inklusion. Dies kann zur Folge haben, daß für  $\underline{x}_k \rightarrow \tau_k$  oder  $\bar{x}_k \rightarrow \tau_{k+p}$  trotz (12) und (16) die Matrix  $[b]$  reelle Matrizen enthält, die singulär oder nicht totalnichtnegativ sind. Dies zeigt das folgende Beispiel:

Beispiel: Es werde  $p = 3$  und  $n = 2$  gewählt. Dann ist im Falle  $\tau_i < [x_k] < \tau_{i+1}$  nach (13) und (14)

$$[b_{i-2,k}] = \frac{[(\tau_{i+1} - \bar{x}_k)^2, (\tau_{i+1} - \underline{x}_k)^2]}{(\tau_{i+1} - \tau_{i-1})(\tau_{i+1} - \tau_i)},$$

$$[b_{i-1,k}] = \frac{[(\underline{x}_k - \tau_{i-1})(\tau_{i+1} - \bar{x}_k), (\bar{x}_k - \tau_{i-1})(\tau_{i+1} - \underline{x}_k)]}{(\tau_{i+1} - \tau_{i-1})(\tau_{i+1} - \tau_i)} + \frac{[(\tau_{i+2} - \bar{x}_k)(\underline{x}_k - \tau_i), (\tau_{i+2} - \underline{x}_k)(\bar{x}_k - \tau_i)]}{(\tau_{i+2} - \tau_i)(\tau_{i+1} - \tau_i)},$$

$$[b_{i,k}] = \frac{[(\underline{x}_k - \tau_i)^2, (\bar{x}_k - \tau_i)^2]}{(\tau_{i+2} - \tau_i)(\tau_{i+1} - \tau_i)}.$$

Wählt man nun

$$\begin{aligned} \tau_i &:= i-1, \quad i = 1(1)5, \\ [x_1] &:= [1+\epsilon, 2-\epsilon] \quad \text{mit } \epsilon \in ]0, 0.5[, \\ [x_5] &:= 2.5, \end{aligned}$$

dann ergibt sich mit  $[b_{ij}] = [\underline{b}_{ij}, \bar{b}_{ij}]$

$$\underline{b}_{11} \underline{b}_{22} - \bar{b}_{12} \bar{b}_{21} = 0.75 (1+\epsilon)\epsilon - (1-\epsilon)^2/16 =: \varphi(\epsilon).$$

Es ist für  $\epsilon \in ]0, 0.5[$

$$\varphi(\epsilon) = 0 \Leftrightarrow \epsilon = (-7 + 2^{1/15})/11 \approx 0.0678 .$$

Für  $\epsilon \leq 0.067$  enthält  $[b]$  reelle  $2 \times 2$ -Matrizen, die singulär oder nicht totalnichtnegativ sind.

In der Praxis werden jedoch die Intervalle  $[x_k]$  zumeist eine so kleine Spanne besitzen, daß  $[b]$  nur totalnichtnegative, reguläre Matrizen enthält. Gleichwohl werden Kriterien benötigt, mit deren Hilfe man entscheiden kann, ob dies der Fall ist. Da diese Fragestellung von allgemeinerem Interesse ist, werden wir sie in einem eigenen Abschnitt, in Abschnitt 4.4, näher untersuchen; denn totalnichtnegative Matrizen treten noch bei einer Reihe weiterer Interpolations- und Approximationsprozesse auf, so z.B. bei der Interpolation mit Funktionen (2.3) (vgl. [20], S. 88) und bei der Approximation mit Splines (vgl. [11], S. 262). Im nächsten Abschnitt werden wir dann auch die Auflösbarkeit und Auflösung des Intervallgleichungssystems für die Bestimmung der Einschließungsintervalle  $[a_j]$ ,

$$\sum_j [b_{ij}] a_j = [y_i] , \quad i = 1(1)n ,$$

in einem allgemeineren Rahmen diskutieren.

#### 4.3.1 Sonderfall: Periodische Splines mit äquidistanten Knoten

In Anlehnung an [11], S. 325-328, fassen wir zunächst einige vom Reellen her bekannte Ergebnisse zusammen.

Es sei eine zweifach unendliche Knotenfolge  $t = (t_i)_{i \in \mathbb{Z}}$  mit

$$t_i := t_0 + ih, \quad i \in \mathbb{Z},$$

und

$$h := \frac{1}{n}$$

gegeben. Mit  $\mathcal{S}_{p,t}^1$  sei die Menge aller 1-periodischen Splines der Ordnung  $p$  mit der Knotenfolge  $t$  bezeichnet.

Eine Funktion  $f \in \mathcal{S}_{p,t}^1$  ist genau dann 1-periodisch, wenn ihre B-Spline-Koeffizientenfolge  $n$ -periodisch ist, d.h. wenn

$$f = \sum_{i \in \mathbb{Z}} \sigma_i b_{i,k,t} \quad \text{mit} \quad \sigma_i = \sigma_{n+i}, \quad i \in \mathbb{Z}.$$

Gegeben seien nun eine Folge von Stützstellen

$$x_k := x_0 + kh, \quad k \in \mathbb{Z},$$

$$\text{mit } x_0 \in ]t_0, t_p[$$

und eine Folge von Stützwerten  $y_k, k \in \mathbb{Z}$ , mit

$y_k = y_{k+nr}$ . Gesucht ist eine Funktion  $f \in \mathcal{S}_{p,t}^1$  mit  $f(x_k) = y_k, k \in \mathbb{Z}$ .

Die Bestimmung der B-Spline-Koeffizienten von  $f$  führt auf das folgende lineare Gleichungssystem

$$c \cdot a = b,$$

$$\text{wobei } a = (\alpha_1, \dots, \alpha_n)^t, \quad b = (y_1, \dots, y_n)^t$$

$$\text{und } c = (c_{jk})_{j,k=1}^n$$

$$\text{mit } c_{jk} := \sum_{r \in \mathbb{Z}} b_{k+rn}(x_0 + jh) = \sum_{r \in \mathbb{Z}} b_n(x_0 + j-k-rn) \quad (17)$$

ist.

Wir wollen nun  $2p-1 \leq n$  voraussetzen. Dann ist höchstens je einer der Summanden in den beiden Summen (17) von Null verschieden.

Die Matrix  $c$  ist von zyklischer Bauart <sup>1)</sup>, d.h. es ist

$$c_{jk} = c_{rs} \text{ , falls } j-k=r-s \pmod{n} \text{ ist,}$$

bzw.  $c$  hat die Gestalt

$$c = \begin{pmatrix} \zeta_1 & \zeta_2 & \zeta_3 & \dots & \zeta_n \\ \zeta_n & \zeta_1 & \zeta_2 & \dots & \zeta_{n-1} \\ \zeta_{n-1} & \zeta_n & \zeta_1 & \dots & \zeta_{n-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \zeta_2 & \zeta_3 & \zeta_4 & \dots & \zeta_1 \end{pmatrix} .$$

Es sei

$$q(x) := \sum_{u=1}^n \zeta_u x^{u-1} , \quad (18)$$

Dann sind mit  $w$  aus (3.5) die Eigenwerte von  $c$  gegeben durch

$$q(\bar{w}^j) , \quad j = 0(1)n-1 , \quad (19)$$

und  $\bar{w}$  mit  $w$  aus (3.9) ist die Modalmatrix zu  $c$  .

Ferner gilt

$$a = \frac{1}{n} \bar{w}^t r w b \quad (20)$$

$$\text{mit } r := \text{diag}(q(1), q(\bar{w}), \dots, q(\bar{w}^{n-1}))^{-1} .$$

Es ist

$$(q(1), q(\bar{w}), \dots, q(\bar{w}^{n-1}))^t = \bar{w}^t (\zeta_1, \dots, \zeta_n)^t . \quad (21)$$

---

1) In der angelsächsischen Literatur wird eine derartige Matrix circulant (matrix) genannt.

Wir beschäftigen uns nun mit dem Intervallinterpolationsproblem vom Typ I. Es seien dazu Intervalle  $[y_k]$  mit  $y_k \in [y_k]$ ,  $k=1(1)n$ , gegeben. Gesucht sind Intervalle  $[c_k]$ ,  $k=1(1)n$ , die die B-Spline-Koeffizienten  $c_k$  der interpolierenden 1-periodischen Spline-Funktion  $f$  enthalten.

Zur Rundungsfehlerkontrolle werden die Werte  $c_{jk}$  mit einer Maschinenintervallarithmetik (vgl. Unterabschn. 3.4.1) entsprechend (13) und (14) berechnet<sup>1)</sup>. Man erhält dann Intervalle  $[c_{jk}]$  mit  $c_{jk} \in [c_{jk}]$ . Die "gerundete" Matrix  $[c] = ([c_{jk}])$  ist i.a. nicht von zyklischer Bauart; da jedoch vom reellen Problem her bekannt ist, daß die Matrix  $c \in [c]$  von zyklischer Bauart ist, liegt  $c$  auch in derjenigen Matrix, die man aus  $[c]$  erhält, indem man die Koeffizienten  $[c_{jk}]$  und  $[c_{rs}]$  mit  $j-k=r-s \pmod{n}$  miteinander schneidet. Die so gebildete Matrix ist von zyklischer Bauart über  $\mathbb{II}(\mathbb{R})$  und sei wieder mit  $[c]$  bezeichnet. Wieder aus dem Grunde, daß vom reellen Problem her bekannt ist, daß die Matrix  $c$  von zyklischer Bauart ist, genügt es, sich im folgendem auf die Teilmenge von  $[c]$

$$\{c\}_Z := \{c' \in M_{n,n}(\mathbb{R}) \mid c' \in [c] \text{ und } c' \text{ ist von} \\ \text{zyklischer Bauart}\}$$

zu beschränken. Da in (18) jeder der Koeffizienten  $c_u$  nur einmal auftritt, läßt sich sofort aus (19) (mit der Optimalitätsaussage in [37], S. 11) ein notwendiges und hinreichendes Kriterium für die Regularität der in  $\{c\}_Z$  enthaltenen reellen Matrizen herleiten:

---

1) Auf die Kennzeichnung von Maschinenintervallen (etwa mit der Tilde wie in Abschn. 3.4) verzichten wir hier.

Es sei

$$\Delta := \{\det c' \mid c' \in \{c\}_Z\} ,$$

$$[q](x) := \sum_{u=1}^n [c_{1u}] x^{u-1} ;$$

dann gilt

$$0 \notin \Delta \Leftrightarrow 0 \notin [q](\bar{w}^{(j)}) , \quad j=0(1)n-1 .$$

Wir setzen nun  $0 \notin \Delta$  voraus. Die Darstellung von  $a$  in (20) legt es für den Fall  $n=2^m$  nahe, den Vektor  $[v] \in II(\mathbb{R}^n)$  'schnell' in der komplexen Intervallarithmetik zu transformieren, um eine Einschließung  $([c_1], \dots, [c_n])^t$  für die B-Spline-Koeffizienten  $c_j$  zu erhalten. Dabei geht  $r$  über in

$$[r] := \text{diag}\left(\frac{1}{[q](1)} , \frac{1}{[q](\bar{w})} , \dots , \frac{1}{[q](\bar{w}^{n-1})}\right) . \quad (22)$$

Es sind dann im wesentlichen 3 FFT, darunter 1 FFT für die Berechnung von  $[r]$  entsprechend (21), durchzuführen, insgesamt werden also nur  $O(nm)$  arithmetische Operationen benötigt im Gegensatz zu  $\frac{2}{3} n^3 + O(n^2)$  Operationen bei Lösung des entsprechenden linearen Gleichungssystems. Da  $[v]$  und  $[r]$  reelle Intervallvektoren sind, kann der Aufwand an arithmetischen Operationen noch etwa halbiert werden, wenn die Vorgehensweise für das entsprechende reelle Problem aus [19] , Abschn. 3.2, auf der Grundlage unserer Ergebnisse aus Abschnitt 3.3 übertragen wird.

#### 4.4 Totalnichtnegative Intervallmatrizen

Wir stellen diesem Abschnitt eine Reihe von Definitionen und Ergebnisse aus der Theorie der totalnichtnegativen Matrizen voran.

##### 4.4.1 Definitionen, Bezeichnungen und Hilfsmittel

Wie üblich (vgl. etwa [16], [35]) bezeichnen wir die Untermatrizen und linearen  $k$ -ter Ordnung einer Matrix  $a \in M_{n,n}(\mathbb{C})$  auf die folgende Weise

$$a \begin{bmatrix} i_1, \dots, i_k \\ j_1, \dots, j_k \end{bmatrix} := \begin{pmatrix} a_{i_1, j_1} & \dots & a_{i_1, j_k} \\ a_{i_2, j_1} & \dots & a_{i_2, j_k} \\ \vdots & & \vdots \\ a_{i_k, j_1} & \dots & a_{i_k, j_k} \end{pmatrix}, \quad (23)$$

$$a \begin{pmatrix} i_1, \dots, i_k \\ j_1, \dots, j_k \end{pmatrix} := \det a \begin{bmatrix} i_1, \dots, i_k \\ j_1, \dots, j_k \end{bmatrix}$$

$$1 \leq i_1 < i_2 < \dots < i_k \leq n, \quad k=1(1)n,$$

$$1 \leq j_1 < j_2 < \dots < j_k \leq n,$$

Unter den führenden Hauptminoren von  $a$  verstehen wir die Minoren

$$a \begin{pmatrix} 1, \dots, k \\ 1, \dots, k \end{pmatrix}, \quad k = 1(1)n.$$

Die Matrix  $a \in M_{n,n}(\mathbb{R})$  heißt totalpositiv bzw. totalnichtnegativ, falls gilt

$$a \begin{pmatrix} i_1, \dots, i_k \\ j_1, \dots, j_k \end{pmatrix} > 0 \quad \text{bzw.} \geq 0 \quad (24)$$

für  $1 \leq i_1 < i_2 < \dots < i_k \leq n$ ,  $k=1(1)n$  .  
 $1 \leq j_1 < j_2 < \dots < j_k \leq n$  ,

Eine totalnichtnegative Matrix  $a \in M_{n,n}(\mathbb{R})$  heißt Oszillationsmatrix, falls es eine positive ganze Zahl  $m$  gibt, derart daß  $a^m$  totalpositiv ist. Oszillationsmatrizen treten beispielsweise in der Mechanik auf (vgl. [20]).

Zum Nachweis der Totalpositivität einer Matrix braucht (24) nur für die Untermatrizen mit aufeinanderfolgenden Zeilen und Spalten nachgewiesen zu werden (vgl. [20], S.299, und [29], S. 85):

Lemma 4.4: Ist  $a \in M_{n,n}(\mathbb{R})$  und gilt

$$a \begin{pmatrix} i, \dots, i+k \\ j, \dots, j+k \end{pmatrix} > 0, \quad i, j = 1(1)n-k, \quad k=0(1)n-1,$$

so ist  $a$  totalpositiv.

Ein schwächeres Kriterium gilt für totalnichtnegative Matrizen (vgl. [16]):

Lemma 4.5: Die Matrix  $a \in M_{n,n}(\mathbb{R})$  besitzt den Rang  $r$ . Dann ist  $a$  genau dann totalnichtnegativ, wenn jeder Minor

$$\alpha \begin{pmatrix} i_1, \dots, i_k \\ j_1, \dots, j_k \end{pmatrix} \text{ mit } \sum_{s=2}^k (j_s - j_{s-1} - 1) \leq n-r$$

nichtnegativ ist für  $k = 1(1)n$ .

Es gilt das folgende Oszillationskriterium (vgl. [20], S. 115):

Lemma 4.1: Eine totalnichtnegative Matrix  $A = (a_{ij}) \in M_{n,n}(\mathbb{R})$  ist genau dann eine Oszillationsmatrix, wenn sie regulär ist und wenn  $0 < a_{ij}$  für  $|i-j| = 1$ ,  $i, j = 1(1)n$ , gilt.

Der Kriterium für den Nachweis der Totalnichtenegativität von Bandmatrizen findet man in [35]. Im Spezialfall, daß ein gegebenes Polynom  $p$  eine Totdiagonalmatrix ist, genügt es für den Nachweis, daß  $A$  totalnichtenegativ und regulär ist, um zeigen, daß die Elemente der beiden Nebendiagonalen nichtnegativ und daß die Hauptminoren sämtlich positiv sind ([20], S. 93-94).

Wir benötigen noch das folgende Lemma (vgl. [20], S. 103, oder [29], S. 89):

Lemma 4.2: Eine totalnichtnegative, reguläre Matrix besitzt nur positive führende Hauptminoren.

Für  $A, B \in M_{n,m}(\mathbb{R})$  definieren wir zwei Transformationen  $a^*, a$  &  $b \in M_{n,m}(\mathbb{R})$  durch

$$(a^*)_{ij} := (-1)^{i+j} a_{ij},$$

$$(a \& b)_{ij} := \begin{cases} a_{ij} \\ b_{ij} \end{cases} \text{ falls } i+j \begin{cases} \text{gerade} \\ \text{ungerade} \end{cases} \text{ ist,} \\ i = 1(1)n, j = 1(1)m.$$

Die Abbildung  $*$  wird häufig "Schachbrett-Transformation" genannt.

Lemma 4.8: Die Abbildung  $*$  ist linear, und es gilt

$$a^{**} = a \quad \text{für alle } a \in M_{n,m}(\mathbb{C}), \quad (25)$$

$$\det a^* = \det a \quad \text{für alle } a \in M_{n,n}(\mathbb{R}), \quad (26)$$

$$a^{*-1} = a^{-1*} \quad \text{für alle regulären } a \in M_{n,n}(\mathbb{C}); \quad (27)$$

die Abbildung  $*$  bildet  $[a] = [\underline{a}, \overline{a}] \in M_{n,n}(\text{II}(\mathbb{R}))$  eineindeutig auf  $[(\underline{a} \& \overline{a})^*, (\overline{a} \& \underline{a})^*]$  ab.

Der Beweis ergibt sich leicht durch Nachrechnen; (26) gilt wegen

$$a^* = r a r$$

$$\text{mit } r := \text{diag}(1, -1, 1, \dots, (-1)^{n+1});$$

bzgl. (27) sei auf [20], S. 86, verwiesen. ▲

Entsprechend definiert man  $[a]^* \in M_{n,m}(\text{II}(\mathbb{C}))$  für  $[a] \in M_{n,m}(\text{II}(\mathbb{C}))$  durch

$$([a]^*)_{ij} := (-1)^{i+j} [a_{ij}], \quad i=1(1)n, j=1(1)m.$$

Mit Lemma 4.8 ergibt sich, daß

$$\{a^* \mid a \in [a]\} = [a]^*$$

gilt, die Wertemenge der Abbildung  $*$  auf  $[a]$  stimmt mit  $[a]^*$  überein. Die Verwendung eines gemeinsamen Symbols für beide Abbildungen ist somit gerechtfertigt.

Wir machen noch darauf aufmerksam, daß durch

$$a \wedge b := (-1)^{i+j} a_{ij} \wedge (-1)^{i+j} b_{ij}, \\ i = 1(1)n, j = 1(1)m, \\ a = (a_{ij}), b = (b_{ij}) \in M_{n,m}(R)$$

auf  $M_{n,m}(R)$  eine Halbordnung herleitet. Die nachstehenden Ergebnisse lassen sich auch für Intervallmatrizen bzgl. dieser Halbordnung formulieren. Da wir dann jedoch zwei Halbordnungen,  $\leq$  und  $\leq$ , parallelisierbar verwenden und daher auch Intervalle bzgl. dieser Halbordnungen verschieden kennzeichnen müßten, verzichten wir hier darauf.

Als neues neue Intervallmatrix  $[a] \in M_{n,n}(I(R))$  (regulär, totalnichteigentlich, invertierbar, totalnichteigentlich, invertierbar) heißt  $[a]$  die entsprechende Eigenschaft besitzen.

(1) Folgendermaßen, falls nicht anders angegeben, ist, stets  $[a] \in M_{n,n}(I(R))$  sein.

Definition 1.1. für den Dispositivität bzw. nicht Dispositivität einer Intervallmatrix

Wir betrachten die spezielle Matrix

$$[a] = [\underline{a}, \bar{a}] := \begin{pmatrix} [1,0] & [1,2] \\ [1,3] & [2,4] \end{pmatrix}.$$

Die beiden zugehörigen Matrizen  $\underline{a}$  und  $\bar{a}$  sind totalnichteigentlich, d. h. sogar totalpositiv, jedoch ist beispielsweise

$$\underline{a} \ \& \ \overline{a} = \begin{pmatrix} 2 & 3 \\ 3 & 2 \end{pmatrix}$$

nicht totalnichtnegativ. Dieses Beispiel zeigt, daß von der Totalpositivität bzw. -nichtnegativität von  $\underline{a}$  und  $\overline{a}$  nicht auf die Totalpositivität bzw. -nichtnegativität von  $a$ ,  $\underline{a} \leq a \leq \overline{a}$ , geschlossen werden darf.

Dieser Schluß ist jedoch möglich, wenn zwei andere Eckmatrizen von  $[a]$  als  $\underline{a}$  und  $\overline{a}$  ausgewählt werden. Dies wollen wir in diesem Unterabschnitt zeigen.

Lemma 4.9: Es sei  $[a] \in M_{n,n}(II(p))$  und es seien  $\underline{a} \ \& \ \overline{a}$ ,  $\overline{a} \ \& \ \underline{a}$  totalnichtnegativ und regulär; dann sind alle  $a^*$  für  $a \in [a]$  invers-isoton, insbesondere ist  $[a]$  regulär.

Beweis: Da  $\underline{a} \ \& \ \overline{a}$  und  $\overline{a} \ \& \ \underline{a}$  totalnichtnegativ und regulär sind, folgert man mittels (27) und der Formel für die Inverse einer Matrix mittels der adjungierten Matrix die Invers-Isotonie von  $(\underline{a} \ \& \ \overline{a})^*$  und  $(\overline{a} \ \& \ \underline{a})^*$ . Dann sind aber nach Lemma 4.1 (b) und Lemma 4.8 auch alle  $a^*$  mit  $a \in [a]$  invers-isoton. ▲

Satz 4.3: Die Intervallmatrix  $[a]$  ist genau dann totalpositiv, wenn  $\underline{a} \ \& \ \overline{a}$  und  $\overline{a} \ \& \ \underline{a}$  totalpositiv sind.

Beweis: Die Bedingung ist offensichtlich notwendig. Der Nachweis, daß sie auch hinreichend ist, wird induktiv erbracht:

Die Behauptung ist im Fall  $n=1$  trivial erfüllt.

Es werde nun angenommen, daß die Behauptung bereits für  $v = 1(1)n-1$ ,  $n \geq 2$ , gezeigt sei.

Es seien  $\underline{a}$ ,  $\bar{a}$   $n \times n$ -Matrizen mit  $\underline{a} \leq \bar{a}$  und es seien  $\underline{a}$  &  $\bar{a}$  und  $\bar{a}$  &  $\underline{a}$  totalpositiv.

Es werde  $a \in [a]$  angenommen. Zum Nachweis der Totalpositivität von  $a$  betrachten wir zunächst die Untermatrizen höchstens  $(n-1)$ -ter Ordnung von  $a$ . Wegen Lemma 4.4 können wir uns dabei auf diejenigen Untermatrizen von  $a$ , die aus aufeinanderfolgenden Zeilen und Spalten gebildet sind, beschränken. Auf diese Untermatrizen und die entsprechenden Untermatrizen von  $\underline{a}$  und  $\bar{a}$  kann die Induktionsvoraussetzung angewendet werden, da die aus diesen Untermatrizen von  $\underline{a}$  und  $\bar{a}$  gebildeten Matrizen

$$\underline{a} \begin{bmatrix} i, \dots, i+k \\ j, \dots, j+k \end{bmatrix} \quad \& \quad \bar{a} \begin{bmatrix} i, \dots, i+k \\ j, \dots, j+k \end{bmatrix}$$

bzw.

$$\bar{a} \begin{bmatrix} i, \dots, i+k \\ j, \dots, j+k \end{bmatrix} \quad \& \quad \underline{a} \begin{bmatrix} i, \dots, i+k \\ j, \dots, j+k \end{bmatrix},$$

als Untermatrizen von  $\underline{a}$  &  $\bar{a}$  bzw.  $\bar{a}$  &  $\underline{a}$  totalpositiv sind.

Es ist noch  $0 < \det a$  zu zeigen. Dies folgt aber wegen der Stetigkeit der Determinanten-Funktion aus  $\det(\underline{a} \& \bar{a})$ ,  $\det(\bar{a} \& \underline{a}) > 0$  und Lemma 4.9. ▲

Bemerkung: Ein entsprechender Satz gilt auch für die allgemeinere Klasse der streng zeichenfesten Intervallmatrizen. Dabei heißt eine reelle Matrix streng zeichenfest, wenn alle Minoren derselben Ordnung ein festes von Null verschiedenes Vorzeichen besitzen (s. [20], S. 284). Wir nennen eine Intervallmatrix streng zeichenfest, wenn sämtliche in ihr

enthaltenen Punktmatrizen streng zeichenfest sind.

Um zu zeigen, daß eine zu Satz 4.3 analoge Äquivalenz auch für die totalnichtnegativen Intervallmatrizen gilt, beweisen wir nachstehend eine leichte Modifikation eines Theorems von Whitney [54] .

Satz 4.4: Es sei  $a \in M_{n,n}(\mathbb{R})$  totalnichtnegativ und es sei  $\epsilon > 0$  . Dann gibt es eine totalpositive Matrix  $c$  mit

$$0 < c_{ij} - a_{ij} < \epsilon, \quad i, j = 1(1)n.$$

Beweis: Die Behauptung ist im Fall  $n=1$  trivial erfüllt, und es sei fortan  $n \geq 2$  gewählt.

Wir setzen

$$f(\sigma) := (e^{-\sigma(i-j)})_{i,j=1}^n;$$

dann ist  $f(\sigma)$  totalpositiv für  $\sigma > 0$  (vgl. [20], S. 89), und es gilt

$$f(\sigma) \rightarrow I \text{ (Einheitsmatrix) für } \sigma \rightarrow \infty. \quad (28)$$

Setzt man

$$\hat{a}(\sigma) := f(\sigma) a f(\sigma), \quad \hat{a}_{ij}(\sigma) = (\hat{a}_{ij}(\sigma)),$$

dann gilt wegen (28)

$$\hat{a}(\sigma) \rightarrow a \quad \text{für } \sigma \rightarrow \infty.$$

Es ist

$$\hat{a}_{ij}(\sigma) = \sum_{u,v=1}^n e^{-\sigma(i-u)} a_{uv} e^{-\sigma(v-j)}, \quad i, j = 1(1)n, \quad (29)$$

und damit

$$a \leq \hat{a}(\sigma). \quad (30)$$

Es gilt  $a < \hat{a}(\sigma)$ , falls  $a$  mindestens zwei nichtverschwindende Koeffizienten besitzt.

Wir wollen nun annehmen, daß  $a$  den Rang  $r$  besitzt. Nach der Entwicklungsformel von Cauchy-Binet (s. [26], S. 12) gilt für  $p = 1(1)n$ :

$$\begin{aligned} \hat{a}(\sigma) \begin{pmatrix} i_1, \dots, i_p \\ k_1, \dots, k_p \end{pmatrix} &= \\ &= \sum \gamma(\sigma) \begin{pmatrix} i_1, \dots, i_p \\ \alpha_1, \dots, \alpha_p \end{pmatrix} a \begin{pmatrix} \alpha_1, \dots, \alpha_p \\ \beta_1, \dots, \beta_p \end{pmatrix} \gamma(\sigma) \begin{pmatrix} \beta_1, \dots, \beta_p \\ k_1, \dots, k_p \end{pmatrix}, \quad (31) \end{aligned}$$

wobei die Summation über alle möglichen Kombinationen  $(\alpha_1, \dots, \alpha_p)$  bzw.  $(\beta_1, \dots, \beta_p)$  von je  $p$  verschiedenen der  $n$  Indizes  $1, 2, \dots, n$  zu erstrecken ist.

In der Matrix  $a$  gibt es für jede Ordnung  $p \leq r$  Minoren, die von Null verschieden sind, wobei alle diese Minoren positives Vorzeichen besitzen. Wegen der Totalpositivität von  $\gamma(\sigma)$  folgt aus (31), daß sämtliche Minoren von  $\hat{a}(\sigma)$  der Ordnung  $p \leq r$  positiv sind. Andererseits sind alle Minoren von  $a$  mit Ordnungen  $> r$  gleich Null,  $\hat{a}(\sigma)$  besitzt also den Rang  $r$ .

Ist  $r = n$ , so sind insbesondere mindestens zwei Koeffizienten von  $a$  von Null verschieden, und wir können  $\sigma_0$  so wählen, daß  $c := \hat{a}(\sigma_0)$  die Behauptung des Satzes erfüllt.

Ist nun  $r < n$ , so bilden wir für  $\delta > 0$  die Matrix  $a^{(1)}(\sigma) = (a_{ij}^{(1)}(\sigma))$ ,

$$a^{(1)}(\sigma) := \hat{a}(\sigma) + \delta e_1 e_1^t,$$

wobei  $e_1 := (1, 0, \dots, 0)^t$  ist.

Es ist  $a^{(1)}(\sigma)$  totalnichtnegativ, und sämtliche Minoren von  $a^{(1)}(\sigma)$  sind bis zur Ordnung  $r$  positiv.  $a^{(1)}(\sigma)$  besitzt sogar den Rang  $r+1$ , da jede Untermatrix  $(r+1)$ -ter Ordnung, die  $a_{11}^{(1)}(\sigma)$  als Koeffizienten enthält, eine positive Determinante besitzt.

Bildet man nun  $\hat{a}^{(1)} = (\hat{a}_{ij}^{(1)})$ ,

$$\hat{a}^{(1)}(\sigma) := f(\sigma) a^{(1)}(\sigma) f(\sigma),$$

dann gilt analog den obigen Überlegungen

$$a_{11}^{(1)}(\sigma) \leq \hat{a}_{11}^{(1)}(\sigma),$$

$$a_{ij}^{(1)}(\sigma) < \hat{a}_{ij}^{(1)}(\sigma), \quad i, j = 1(1)n, \quad i \cdot j \neq 1,$$

und damit wegen  $\hat{a}_{11}(\sigma) < a_{11}^{(1)}(\sigma)$  und (30) auch

$$a < \hat{a}^{(1)}(\sigma);$$

weiter gilt:  $\hat{a}^{(1)}(\sigma)$  besitzt den Rang  $r+1$ .

Mit

$$a^{(2)}(\sigma) := \hat{a}^{(1)}(\sigma) + \delta e_1 e_1^t$$

erhält man dann eine totalnichtnegative Matrix vom Rang  $r+2$ .

Führt man mit dieser Konstruktion entsprechend weiter fort, so erhält man nach  $n-r$  vielen Schritten die Matrix  $\hat{a}^{(n-r)}(\sigma)$ , die totalpositiv ist für  $\sigma$ ,  $\delta > 0$  und für die gilt:

$$\hat{a}^{(n-r)}(\sigma) =$$

$$= f(\sigma)(f(\sigma)(\dots f(\sigma)(f(\sigma)a f(\sigma) + \delta e_1 e_1^t) f(\sigma) + \delta e_1 e_1^t) f(\sigma) + \dots + \delta e_1 e_1^t) f(\sigma)$$

$$= f(\sigma)^{n-r+1} a f(\sigma)^{n-r+1} + \sum_{u=1}^{n-r} f(\sigma)^u \delta e_1 e_1^t f(\sigma)^u = a .$$

Für genügend großes  $\sigma$  und genügend kleines  $\delta$  besitzt daher  $\hat{a}^{(n-r)}(\sigma)$  die in der Behauptung des Satzes verlangte Approximationseigenschaft.  $\Delta$

Satz 4.5: Die Intervallmatrix  $[a]$  ist genau dann totalnichtnegativ, wenn  $\underline{a} \approx \bar{a}$  und  $\bar{a} \approx \underline{a}$  totalnichtnegativ sind.

Beweis: Es genügt wieder nur zu zeigen, daß die Bedingung des Satzes hinreichend ist. Zur Abkürzung setzen wir

$$\begin{aligned} b &:= \underline{a} \approx \bar{a} , \quad c := \bar{a} \approx \underline{a} , \\ \Gamma &:= \{(i,j) \in \{1,\dots,n\}^2 \mid \underline{a}_{ij} = \bar{a}_{ij}\} , \\ \Gamma' &:= \{1,\dots,n\}^2 \setminus \Gamma . \end{aligned}$$

O.B.d.A. nehmen wir an, daß  $n \geq 2$ ,  $\emptyset \neq \Gamma$ ,  $\Gamma'$  ist. Nach Satz 4.4 gibt es zwei Folgen totalpositiver Matrizen  $\{u^{(v)}\}$ ,  $\{v^{(v)}\}$  mit

$$\left. \begin{aligned} \lim_{v \rightarrow \infty} u^{(v)} &= b , \quad b < u^{(v)} \quad \text{für alle } v , \\ \lim_{v \rightarrow \infty} v^{(v)} &= c , \quad c < v^{(v)} \quad \text{für alle } v . \end{aligned} \right\} \quad (32)$$

Wir zeigen nun, daß diese Folgen Teilfolgen  $\{u^{(v_\lambda)}\}$ ,  $\{v^{(v_\lambda)}\}$  mit den folgenden Eigenschaften besitzen:

$$\left. \begin{aligned} \underline{a}_{ij} < u_{ij}^{(v_\lambda)} < \bar{a}_{ij} < v_{ij}^{(v_\lambda)} , \quad \text{falls } i+j \text{ gerade} \\ \underline{a}_{ij} < v_{ij}^{(v_\lambda)} < \bar{a}_{ij} < u_{ij}^{(v_\lambda)} , \quad \text{falls } i+j \text{ ungerade} \end{aligned} \right\} \text{ und } \begin{matrix} (i,j) \in \Gamma' \\ \text{ist,} \end{matrix} \quad (33)$$

und

$$(-1)^{i+j} u_{ij}^{(v_\lambda)} \leq (-1)^{i+j} v_{ij}^{(v_\lambda)} , \text{ falls } (i,j) \in \Gamma \text{ ist.} \quad (34)$$

Zur Konstruktion der Teilfolgen  $\{u^{(v_\lambda)}\}$  ,  $\{v^{(v_\lambda)}\}$  :

Ist  $v_0 \in \mathbb{N}$  genügend groß gewählt, dann läßt sich (33) für  $v \geq v_0$  stets erfüllen.

$\Gamma$  werde zerlegt in

$$\Gamma = \Gamma_1 \cup \Gamma_2$$

mit  $\Gamma_1 := \{(i,j) \in \Gamma \mid i+j \text{ gerade}\}$  ,

$\Gamma_2 := \{(i,j) \in \Gamma \mid i+j \text{ ungerade}\}$  .

Wir setzen

$$a_{ij} := \underline{a}_{ij} = \bar{a}_{ij} \quad \text{für } (i,j) \in \Gamma .$$

Zu  $e^{(v_0)}$  ,

$$e^{(v_0)} := \min \left\{ \min_{(i,j) \in \Gamma_1} \{v_{ij}^{(v_0)} - a_{ij}\}, \min_{(i,j) \in \Gamma_2} \{u_{ij}^{(v_0)} - a_{ij}\} \right\} ,$$

existiert ein  $v_1 \in \mathbb{N}$  ,  $v_1 \geq v_0$  , so daß für alle  $v, v \geq v_1$ , gilt

$$\left. \begin{aligned} a_{ij} < u_{ij}^{(v)} &\leq v_{ij}^{(v)} && \text{für } (i,j) \in \Gamma_1 , \\ a_{ij} < v_{ij}^{(v)} &\leq u_{ij}^{(v)} && \text{für } (i,j) \in \Gamma_2 . \end{aligned} \right\} \quad (35)$$

Zusammen mit (33) gilt folglich

$$(-1)^{i+j} u_{ij}^{(v_1)} \leq (-1)^{i+j} v_{ij}^{(v_1)} , \quad i, j = 1(1)n .$$

Nun definiert man

$$e^{(v_1)} := \min \left\{ \min_{(i,j) \in \Gamma_1} \{v_{ij}^{(v_1)} - a_{ij}\}, \min_{(i,j) \in \Gamma_2} \{u_{ij}^{(v_1)} - a_{ij}\} \right\} ,$$

dann gibt es ein  $v_2$  ,  $v_2 \geq v_1$  , welches (35) mit  $v_2$  anstelle von  $v_1$  und  $v_1$  anstelle von  $v_0$  erfüllt. Analog fährt man mit der Konstruktion der Indexfolge  $\{v_\lambda\}$  weiter fort.

Mit  $\{r^{(v)}\}$  bzw.  $\{m^{(v)}\}$  sei die Teilfolge  $\{u^{(v_\lambda)}\}$  bzw.  $\{b^{(v_\lambda)}\}$  bezeichnet. Mit (33) und (34) gilt

$$r^{(v)} \& m^{(v)} \leq m^{(v)} \& r^{(v)} .$$

Wegen

$$(r^{(v)} \& m^{(v)}) \& (m^{(v)} \& r^{(v)}) = r^{(v)} ,$$

$$(m^{(v)} \& r^{(v)}) \& (r^{(v)} \& m^{(v)}) = m^{(v)}$$

folgt mit Anwendung von Satz 4.3 auf das Intervall  $[r^{(v)} \& m^{(v)}, m^{(v)} \& r^{(v)}]$ , daß sämtliche reelle Matrizen aus diesem Intervall totalpositiv sind.

Nach (32) gilt ferner

$$r^{(v)} \& m^{(v)} \rightarrow \underline{a} \text{ und } m^{(v)} \& r^{(v)} \rightarrow \bar{a} \text{ für } v \rightarrow \infty .$$

Zu jedem  $\alpha \in [\underline{a}]$  gibt es eine Folge von reellen Matrizen  $\{\alpha^{(v)}\}$  mit  $\alpha^{(v)} \in [r^{(v)} \& m^{(v)}, m^{(v)} \& r^{(v)}]$ , die gegen  $\alpha$  konvergiert. Wegen der Totalpositivität der Matrizen  $\alpha^{(v)}$  folgt die Totalnichtnegativität von  $\alpha$  .

▲

Mit Lemma 4.9 erhalten wir:

Korollar 4.1: Die Intervallmatrix  $[a]$  ist genau dann regulär und totalnichtnegativ, wenn  $\underline{a} \& \bar{a}$  und  $\bar{a} \& \underline{a}$  regulär und totalnichtnegativ sind.

Hieraus folgt mit Lemma 4.6:

Korollar 4.2: Die Intervallmatrix  $[a] = [\underline{a}, \bar{a}]$  mit  $\underline{a} = (\underline{a}_{ij})$  ist genau dann eine Oszillationsmatrix, wenn  $\underline{a}$  &  $\bar{a}$  und  $\bar{a}$  &  $\underline{a}$  regulär und totalnichtnegativ sind und wenn gilt

$$0 < \underline{a}_{ij} \quad \text{für } |i-j| = 1, i, j = 1(1)n.$$

Satz 4.6: Ist  $[a] = [\underline{a}, \bar{a}]$  totalnichtnegativ und gilt  $\underline{a} < \bar{a}$ , dann sind alle reellen Matrizen  $a \in [a]$  mit  $\underline{a} < a$  totalpositiv.

Beweis: Es sei  $a \in [a]$  mit  $\underline{a} < a$ . Wie im Beweis zu Satz 4.5 findet man totalpositive Matrizen  $u$  und  $v$  mit

$$\begin{aligned} \underline{a}_{ij} < u_{ij} < a_{ij} \leq \bar{a}_{ij} < v_{ij}, & \text{ falls } i+j \text{ gerade ist,} \\ \underline{a}_{ij} < v_{ij} < a_{ij} \leq \bar{a}_{ij} < u_{ij}, & \text{ falls } i+j \text{ ungerade ist.} \end{aligned}$$

Insbesondere ist  $a \in [u \& v, v \& u]$ . Wie im Beweis zu Satz 4.5 folgert man, daß  $a$  totalpositiv ist.  $\blacktriangle$

#### 4.4.3 Anwendung auf lineare Intervallgleichungssysteme

Für lineare Intervallgleichungssysteme mit totalnichtnegativen, regulären  $2 \times 2$ -Intervallmatrizen und rechten Seiten, die gewissen Bedingungen genügen, wurde in [38] die optimale Intervalleinschließung der Lösungsmenge  $\{x\}$  angegeben. Wir wollen nun das Ergebnis von Lemma 4.2 für lineare Intervallgleichungssysteme mit beliebigen totalnichtnegativen, regulären Intervallmatrizen und "oszillierenden" rechten

Seiten fruchtbar machen.<sup>1)</sup> Der folgende Satz zeigt, daß unter ähnlichen Voraussetzungen an die rechte Seite wie in Lemma 4.2 nur zwei reelle lineare Gleichungssysteme aufgelöst zu werden brauchen, um  $\text{Intv}\{r\}$  zu bestimmen.

Satz 4.7: Es sei  $[a]$  totalnichtnegativ und regulär. Ferner sei  $[c] \in \text{II}(\mathbb{R}^n)$  und es gelte einer der folgenden drei Fälle

- (a)  $0 \leq [c]^*$ ,
- (b)  $0 \geq [c]^*$ ,
- (c)  $0 \in [c]$ .

Dann gilt mit

$$\left. \begin{aligned} u &:= (\underline{a} \& \bar{a})^{-1}, \quad v := (\bar{a} \& \underline{a})^{-1}, \\ r &:= \underline{c} \& \bar{c}, \quad l := \bar{c} \& \underline{c}, \end{aligned} \right\} \quad (36)$$

$$\text{Intv}[r] = \begin{cases} [(vr)^*, (ul)^*]^* & \text{im Fall (a),} \\ [(ur)^*, (vl)^*]^* & \text{im Fall (b),} \\ [(ur)^*, (ul)^*]^* & \text{im Fall (c).} \end{cases}$$

Beweis: Wir zeigen zunächst, daß für jede nichtleere, beschränkte Teilmenge  $\Theta$  des  $\mathbb{R}^n$  gilt

$$\text{Intv}\{p^* | p \in \Theta\} = (\text{Intv } \Theta)^*. \quad (37)$$

Es werde die  $i$ -te Komponente des Infimums des links stehenden Intervallvektors betrachtet. Nach Lemma 1.2 gilt ( $e_i$  bezeichne den  $i$ -ten Einheitsvektor)

---

1) Die Ergebnisse des folgenden Satzes lassen sich auch mittels Anwendung des Verfahrens von Kupermann und Hansen (s. [2], Abschnitt 17) herleiten.

$$(\inf \text{Intv}\{p^* | p \in \Theta\})_i = \inf\{(-1)^{i+1} e_i^t p | p \in \Theta\} \\ = (\inf \Theta \& \sup \Theta)_i^* ;$$

dies ist gerade das Infimum der i-ten Komponente des rechts stehenden Intervallvektors.

Ferner gilt

$$a \cdot r = c \Leftrightarrow a^* \cdot r^* = c^* . \quad (38)$$

Wir führen den Beweis der Behauptung des Satzes nur für den Fall (a) - die übrigen Fälle ergeben sich mit analoger Beweisführung - .

Nach Lemma 4.9 ist  $[(\underline{a} \& \bar{a})^*, (\bar{a} \& \underline{a})^*]$  invers-isoton, woraus mit

$$\{y\} := \{y \in \mathbb{R}^n | \exists a \in [a] : a^* y \in [c]^*\} ,$$

(27), Lemma 4.2 und den Bezeichnungen (36) folgt

$$\text{Intv}\{y\} = [(\bar{a} \& \underline{a})^{*-1} \inf([c]^*), (\underline{a} \& \bar{a})^{*-1} \sup([c]^*)] \\ = [v^* r^*, u^* l^*] \\ = [(vr)^*, (ul)^*] .$$

Mit (37) und (38) folgt daraus die Behauptung. A

Analog zeigt man

Satz 4.8: Ist  $[a]$  totalnichtnegativ und regulär, dann ist

$$\text{Intv}\{a^{-1} | a \in [a]\} = [(\bar{a} \& \underline{a})^{-1*}, (\underline{a} \& \bar{a})^{-1*}]^* .$$

Aus Lemma 4.7 folgt nach einem wohlbekannten Satz, daß der reelle Gauß-Algorithmus, angewendet auf totalnientnegative, reguläre reelle Matrizen, ohne Pivotisierung durchführbar ist.

Eingedenk Satz 4.1, kann man nun vermuten, daß der IGA bei totalnichtnegativen, regulären Intervallmatrizen ohne Pivotisierung durchführbar ist. Dies ist jedoch, wie wir im folgenden Beispiel zeigen, nicht der Fall.

Es sei

$$[a] := \begin{pmatrix} [4,5] & [2,3] & 1 \\ [2,3] & 4 & [2,3] \\ 1 & [2,3] & [4,5] \end{pmatrix}. \quad (39)$$

Nach Korollar 4.2 ist  $[a]$  eine Oszillationsmatrix. Der IGA ohne Pivotisierung liefert nach dem ersten Eliminationsschritt die Restmatrix

$$\begin{pmatrix} [\frac{7}{4}, \frac{16}{5}] & [\frac{5}{4}, \frac{13}{5}] \\ [\frac{5}{4}, \frac{13}{5}] & [\frac{15}{4}, \frac{24}{5}] \end{pmatrix}$$

und nach dem zweiten Eliminationsschritt

$$[a_{33}^{(\cdot)}] = [-\frac{79}{700}, \frac{5519}{1280}] ,$$

also

$$0 \in [a_{33}^{(\cdot)}]$$

und bricht ab. Aus dem Beispiel kann man ablesen, daß der IGA sogar auch dann abbrechen kann, wenn Zeilen und Spaltenvertauschungen zugelassen, aber nur dann vorgenommen werden, wenn  $[a_{jj}^{(j-1)}]$  ein Nullintervall ist. Es sei jedoch darauf hingewiesen, daß in diesem Beispiel bei anderer Pivotwahl - etwa  $[a_{11}]$  als erstem Pivot - ein Abbruch vermieden werden kann.

Satz 4.9: Es sei  $[a]$  totalnichtnegativ, regulär und tridiagonal, d.h.

$$0 = [a_{ij}], \text{ falls } 1 < |i-j| \text{ ist, } i, j = 1(1)n.$$

Dann ist der IGA ohne Pivotisierung durchführbar.

Beweis: Es sei  $a \in [a]$ , dann sind nach Lemma 4.7 sämtliche führende Hauptminoren von  $a$  von Null verschieden. Das Nichtverschwinden der führenden Hauptminoren sämtlicher in einer Intervallmatrix enthaltenen reellen Matrizen ist aber nach [44] ein hinreichendes Kriterium für die Durchführbarkeit des IGA, angewendet auf eine tridiagonale Intervallmatrix.

▲

Satz 4.10: Der IGA, angewendet auf die Intervallmatrix  $[a]$ , ist genau dann ohne Pivotisierung durchführbar, wenn dies für  $[a]^*$  der Fall ist.

Beweis: Wendet man den IGA ohne Pivotisierung auf die Matrizen  $[a]$  und  $[a]^*$  an, dann ist in jedem Schritt die Restmatrix von  $[a]^*$  die Schachbrett-transformierte Restmatrix von  $[a]$ :

$$\begin{aligned} (-1)^{i+k} [a_{ik}^{(j)}] - \frac{(-1)^{i+j} [a_{ij}^{(j)}]}{(-1)^{j+j} [a_{jj}^{(j)}]} (-1)^{j+k} [a_{jk}^{(j)}] \\ = (-1)^{i+k} \left( [a_{ik}^{(j)}] - \frac{[a_{ij}^{(j)}]}{[a_{jj}^{(j)}]} [a_{jk}^{(j)}] \right) \\ = (-1)^{i+k} [a_{ik}^{(j+1)}] \end{aligned}$$

$$i, k = j+1(1)n, j = 1(1)n-1.$$

▲

Insbesondere bricht der IGA für die Matrix  $[a]^*$  mit  $[a]$  aus (39) ab, falls Zeilen- und Spaltenvertauschungen nur dann vorgenommen werden, wenn  $[a_{jj}^{(j-1)}]$  ein Nullintervall ist. Jede M-Matrix  $a \in M_{n,n}(\mathbb{R})$  besitzt die Eigenschaft, daß

$$a \begin{bmatrix} i_1, \dots, i_k \\ i_1, \dots, i_k \end{bmatrix} \quad \text{für } 1 \leq i_1 < \dots < i_k \leq n, \quad k = 1(1)n,$$

inversisoton ist (vgl. etwa [34], S. 36). Jede M-Matrix  $a$  hat damit auch die Eigenschaft, daß

$$a \begin{bmatrix} i, \dots, i+k \\ i, \dots, i+k \end{bmatrix} \quad \text{für } 0 \leq k \leq n-i, \quad i = 1(1)n, \quad (40)$$

inversisoton ist.

Sämtliche Punktmatrizen, die in  $[a]^*$  mit  $[a]$  nach (39) enthalten sind, besitzen die Eigenschaft (40) (die Inversisotonie von  $[a]^*$  ergibt sich mit Lemma 4.9, die Inversisotonie der echten Untermatrizen von  $[a]^*$  beweist man leicht durch Nachrechnen). Das Beispiel der Matrix  $[a]^*$  zeigt daher, daß das Ergebnis des Satzes 4.1 für M-Matrix-Intervalle nicht auf die (etwas) allgemeinere Klasse der Intervallmatrizen, deren Punktmatrizen sämtlich die Eigenschaft (40) besitzen, verallgemeinert werden kann.

Wir möchten noch erwähnen, daß sich mit Anwendung von Satz 1.6 [20], S. 125, sofort Schranken für die Eigenwerte einer Matrix aus einem Oszillationsmatrixintervall gewinnen lassen: Nach [20], S. 100, besitzt eine Oszillationsmatrix  $a$  nur positive, paarweise verschiedene Eigenwerte  $\lambda_i(a)$ , die wir

uns folgendermaßen angeordnet denken

$$0 < \lambda_n(a) < \dots < \lambda_2(a) < \lambda_1(a) .$$

Satz 4.11: Es sei  $[a]$  eine Oszillationsmatrix. Dann gilt für die Eigenwerte  $\lambda_i(a)$  jeder reellen Matrix  $a \in [a]$

$$\lambda_n(\underline{a} \ \& \ \overline{a}) \leq \lambda_i(a) \leq \lambda_1(\overline{a}) , \ i = 1(1)n .$$

## Kapitel 5 Intervallinterpolation mit den Funktionen

$$\alpha \xi^{\beta} \text{ und } \alpha_1 \exp(\beta_1 \xi) + \alpha_2 \exp(\beta_2 \xi)$$

In diesem Kapitel beschäftigen wir uns mit zwei Funktionensystemen  $\mathfrak{F}$ , bei denen die (eindeutige) Lösbarkeit der reellen Interpolationsaufgabe nicht gegeben ist, d.h.

zu  $\xi_k, \eta_k$ ,  $k = 1(1)n$ , gibt es nicht notwendig ein (und genau ein)  $\varphi \in \mathfrak{F}$  mit  $\varphi(\xi_k) = \eta_k$ .

Im Gegensatz zu den in Kap. 2 behandelten Funktionensystemen kann hier die optimale Funktionsstreifeneinschließung der Lösungsfamilie höchstens in Sonderfällen angegeben werden. Für die Intervallinterpolation mit den Funktionen  $\alpha \xi^{\beta}$  beweisen wir ein notwendiges und hinreichendes Kriterium dafür, daß die Menge der Parameterpaare  $(\alpha, \beta)$  nicht leer ist. Ist dies der Fall, dann kann die optimale Intervalleinschließung dieser Menge angegeben werden.

Sind die Parameter der Lösungsfamilie eingeschränkt, dann kann bei Hinzunahme weiterer Intervalle  $[\xi_k]$  und  $[\eta_k]$  entweder die bereits erhaltene Einschließung weiter verbessert werden oder es kann sich die Nichtexistenz einer reellen Lösung herausstellen. Diese Verbesserung (s. Abb.4), auf die schon in [39] hingewiesen wurde, wird durch das fundamentale Prinzip, der Intervallrechnung, die Durchschnittsbildung, erreicht. Wir gehen darauf in Abschn. 5.1 näher ein.

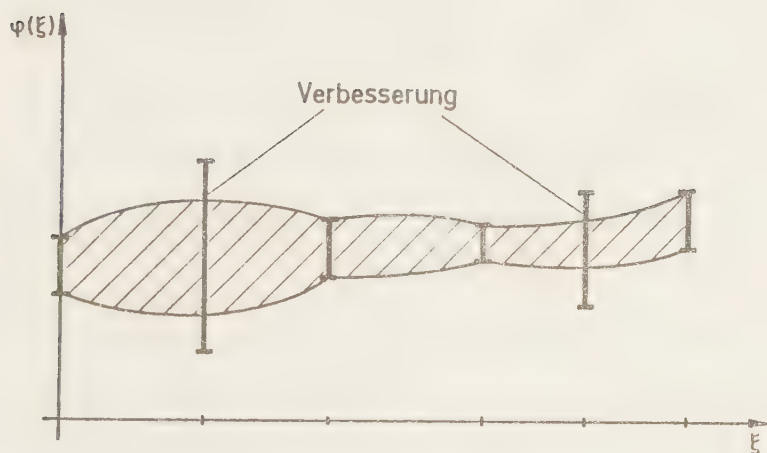


Abb. 4 Funktionsstreifeneinschließung mit  
Verbesserung von Meßintervallen  
(nach [39])

### 5.1 Intervallinterpolation mit den Funktionen $\alpha \xi^B$

In der Biologie haben Wachstumsprozesse beispielsweise eines Tumors (vgl. [7]) oft einen exponentiellen Verlauf. Man wird in derartigen Fällen zur Interpolation die Funktionenklasse

$$\mathfrak{F} := \{ \alpha \xi^B : \mathbb{R}_+ \rightarrow \mathbb{R} \mid \alpha, B \in \mathbb{R} \}$$

verwenden. Gesucht ist im Falle der Intervallinterpolation vom Typ II die Menge  $\mathfrak{M}$  aller Parameterpaare  $(\alpha, B) \in \mathbb{R}^2$ , derart daß gilt

$$\alpha \xi_k^B \in [\eta_k] \text{ für mindestens ein } \xi_k \in [\varepsilon_k], k=1(1)n. \quad (1)$$

Wie man sich leicht klarmacht, kann man sich auf die Fälle

$$\alpha > 0 \text{ und } [\eta_k] > 0, k=1(1)n, \quad (2)$$

beschränken. Die Bedingung (2) soll im folgenden stets vorausgesetzt sein.

Wir wollen zunächst den Fall  $1 \in [\varepsilon_k]$  betrachten, um ihn dann später ausschließen zu können: Aus  $\alpha \xi_k^B \in [\eta_k]$  und  $\varepsilon_k = 1$  folgt  $\alpha \in [\eta_k]$ , der Parameter  $\alpha$  wird also von vornherein auf das Intervall  $[\eta_k]$  eingeschränkt.

Wir zerlegen den Indexbereich in zwei Teilmengen I und J mit  $I \cup J = \{1, \dots, n\}$ , derart daß gilt

$$0 < [\varepsilon_k] < 1 \Leftrightarrow k \in I,$$

$$1 < [\varepsilon_k] \Leftrightarrow k \in J.$$

Wir setzen für  $[\xi] = [\underline{\xi}, \bar{\xi}] \in II(R)$

$$\exp [\xi] := [\exp \underline{\xi}, \exp \bar{\xi}],$$

$$\ln [\xi] := [\ln \underline{\xi}, \ln \bar{\xi}] \text{ für } [\xi] > 0.$$

Dann sind mit den Setzungen

$$\left. \begin{aligned} a &:= \ln \alpha \\ [x_k] &:= \ln [\xi_k], [y_k] := \ln [\eta_k], k=1(1)n, \end{aligned} \right\} (3)$$

die Bedingungen (1) äquivalent zu

$$B \in ([y_k] - a) / [x_k], k=1(1)n.$$

Wir betrachten zunächst den Sonderfall, daß  $a$  bekannt ist;

o.B.d.A. sei  $a = 0$ , d.h.  $\alpha = 1$ , angenommen. Ist unter

dieser Voraussetzung

$$[B] := \bigcap_{k=1}^n [y_k] / [x_k]$$

leer, so ist auch  $\Theta$  leer; andernfalls gilt

$$\Theta = \{1\} \times [B]$$

und

$$\xi[B] := \{\xi^B | B \in [B]\}, \xi \in \mathbb{R}_+,$$

ist die optimale Intervalleinschließung der Lösungsfamilie.

Wir kehren nun zu der allgemeineren Fragestellung (1)

zurück und setzen für  $i, j = 1(1)n, i < j$ ,

$$P_{ij} := \{(a, \beta) \in \mathbb{R}_+ \times \mathbb{R} | \exists x_k \in [x_k] : a + \beta x_k \in [y_k], k=i, j\}. \quad (4)$$

Dies ist gerade die Parametermenge des gemäß (3) transformierten Teilproblems für die Intervalle  $[x_k], [y_k], k=i, j$ .

Die Parametermenge  $P$ ,

$$P := \{(a, \beta) \in \mathbb{R}_+ \times \mathbb{R} \mid \exists x_k \in [x_k] : a + \beta x_k \in [y_k], k=1(1)n\},$$

des transformierten Gesamtproblems ergibt sich dann gerade als der Durchschnitt der Parametermengen sämtlicher Teilprobleme

$$P = \bigcap_{i < j} P_{ij}. \quad (5)$$

Die Menge  $P_{ij}$  besteht aus allen möglichen Koeffizientenpaaren  $(a, \beta)$  derjenigen Polynome 1. Grades  $a + \beta x$ , die auf  $[x_k]$  einen Wert aus  $[y_k]$ ,  $k=i, j$ , annehmen. Es gilt also für  $i, j = 1(1)n$ ,  $i < j$ ,

$$P_{ij} = \left\{ \frac{1}{x_j - x_i} \begin{pmatrix} x_j & -x_i \\ -1 & 1 \end{pmatrix} \begin{pmatrix} y_i \\ y_j \end{pmatrix} \mid x_k \in [x_k], y_k \in [y_k], k=i, j \right\}$$

und damit

$$\text{Intv } P_{ij} \subseteq Q_{ij} := \frac{1}{[x_j] - [x_i]} \begin{pmatrix} [y_i][x_j] - [y_j][x_i] \\ [y_j] - [y_i] \end{pmatrix}. \quad (6)$$

Im Fall, daß die Intervalle  $[x_k]$  sämtlich Punktintervalle sind, gilt in (6) die Gleichheit (s. Abschn. 2.2). Andernfalls gilt i.a. nur die echte Inklusion. Unter der Voraussetzung  $P \neq \emptyset$  erhalten wir mit Lemma 1.3 und (5) die folgende Einschließung für die optimale Intervalleinschließung der Parametermenge  $P$

$$\text{Intv } P \subseteq \bigcap_{i < j} \text{Intv } P_{ij} \subseteq \bigcap_{i < j} Q_{ij}. \quad (7)$$

Der Durchschnitt  $Q$  dieser Einschließungen,

$$Q := \bigcap_{i < j} Q_{ij}, \quad (8)$$

läßt sich darstellen als  $Q = Q_1 \times Q_2$  mit  $Q_1, Q_2 \in II(R)$ .

Ist  $Q$  leer, so ist auch die Parametermenge  $\Theta$  des ursprünglichen Problems leer. Im Fall  $Q \neq \emptyset$  erhalten wir die Einschließung

$$\Theta \subseteq (\exp Q_1) \times Q_2. \quad (9)$$

Diese Einschließung kann mit etwas erhöhten Rechenaufwand noch verschärft werden. Es gilt nämlich, wie wir für allgemeinere monotone Funktionsstreifen zeigen werden, im Fall  $P \neq \emptyset$  unter gewissen Voraussetzungen die Gleichung

$$\text{Intv } P = \cap \text{Intv } P_{ij}. \quad (10)$$

Daß diese Gleichung nicht für beliebige Mengen  $P_{ij}$  mit  $P$ , definiert nach (5), gültig ist, hatten wir schon in Abschnitt 1.2 gezeigt. Wir wenden uns nun der Bestimmung von  $\text{Intv } P_{ij}$  zu und setzen

$$[\varphi_k](a) = [\underline{\varphi}_k(a), \bar{\varphi}_k(a)] := ([y_k] - a) / [x_k], \quad a \in R, \quad k=1(1)n. \quad (11)$$

Dann ist

$$S_{ij} := \{(a, b) \in R_+ \times R \mid ([\varphi_i](a) \cap [\varphi_j](a) \neq \emptyset) \wedge (b \in [\varphi_m]^{-1}(a), m=1, j)\}, \\ i, j=1(1)n \text{ mit } i < j$$

die Schnittfläche der Streifen  $[\varphi_i]$  und  $[\varphi_j]$ .

Aus (4) folgt sofort, daß gilt

$$S_{ij} = P_{ij}, \quad i, j = 1(1)n.$$

Im Fall  $k \in I$  bzw.  $k \in J$  ist  $[\varphi_k]$  streng isoton bzw. streng antiton.

Lemma 5.1: Jede der Randfunktionen des Funktionsstreifen  $[\varphi_i]$  besitzt mit jeder der Randfunktionen des Streifen

$[\varphi_j]$  für  $i, j = 1(1)n$  mit  $i \neq j$  genau einen Schnittpunkt.

Beweis: Der Beweis der Behauptung für den Fall, daß die Intervalle  $[x_k]$  Punktintervalle sind, ist unmittelbar klar; Abszisse und Ordinate des Schnittpunktes sind

$$\frac{1}{x_j - x_i} (y_i x_j - y_j x_i, y_j - x_i),$$

wobei  $y_k \in \{\underline{y}_k, \bar{y}_k\}$ ,  $k=i, j$ , je nachdem welche der Randfunktionen miteinander geschnitten werden.

Die Intervalle  $[x_k]$  seien nun nicht notwendig Punktintervalle. Im Fall  $[x_i] < 0 < [x_j]$  sind die Randfunktionen von  $[\varphi_i]$  streng isoton und von  $[\varphi_j]$  streng antiton und die Behauptung des Satzes ist erfüllt.

Es sei nun  $\text{sign}[x_i] = \text{sign}[x_j]$  und  $\varphi_k \in \{\underline{\varphi}_k, \bar{\varphi}_k\}$ ,  $k=i, j$ . Wir nehmen an, daß  $\varphi_i$  und  $\varphi_j$  zwei Schnittpunkte besitzen; dann existieren  $x_k, x'_k \in \{\underline{x}_k, \bar{x}_k\}$ ,  $y_k \in \{\underline{y}_k, \bar{y}_k\}$ ,  $k=i, j$ , und  $a_1, a_2$  mit

$$a_1 = \frac{y_i x_j - y_j x_i}{x_j - x_i} \leq \max\{y_i, y_j\} =: z,$$

$$a_2 = \frac{y_i x'_j - y_j x'_i}{x'_j - x'_i} \geq \min\{y_i, y_j\} =: z'.$$

Daraus folgt

$$y_i x_j - y_j x_i \leq z x_j - z x_i,$$

$$z' x'_j - z' x'_i \leq y_i x'_j - y_j x'_i.$$

Sowohl für  $z = y_j$  und  $z' = y_i$  als auch für  $z = y_i$  und  $z' = y_j$  ergibt sich wegen  $\text{sign}[x_i] = \text{sign}[x_j]$  sowohl  $y_i \leq y_j$  als auch  $y_j \leq y_i$ , also  $y_j = y_i$ , und damit  $a_1 = a_2$ .

Mit einer ähnlichen Argumentation schließt man, daß beide Funktionen mindestens einen Schnittpunkt besitzen. Damit haben sie genau einen Schnittpunkt.  $\blacktriangle$

Die Gleichung (10) besagt, daß nur die optimalen Intervalleinschließungen der Parametermengen  $P_{ij}$  berechnet werden müssen, um die optimale Intervalleinschließung der (nur in Sonderfällen explizit angebbaren) Menge  $P$  zu bestimmen. Die optimalen Intervalleinschließungen der Mengen  $P_{ij}$  lassen sich nach Lemma 5.1 leicht aus den Schnittpunkten der Randfunktionen von  $[q_k]$  berechnen. Welche Schnittpunkte dabei jeweils in Frage kommen, wird sich aus den nachfolgenden Untersuchungen ergeben.

Mit

$$\left. \begin{aligned} [a] &:= \exp(\pi_1(\text{Intv } P)), \\ [B] &:= \pi_2(\text{Intv } P) \end{aligned} \right\} \quad (12)$$

ist

$$[a]\xi^{[B]} := \{ \alpha \xi^B \mid \alpha \in [a], B \in [B] \}, \xi \in \mathbb{R}_+,$$

eine Funktionsstreifeneinschließung der Lösungsfamilie. Wir deuten nun die Intervalle  $[\eta_k]$  wie in Abschnitt 1.1 als Meßintervalle.

Es sei

$$[a][\xi_k]^{[B]} := \{ \alpha \xi_k^B \mid \alpha_k \in [\xi_k], \alpha \in [a], B \in [B] \}, k=1(1)n.$$

Gilt dann für ein  $j \in \{1, \dots, n\}$

$$[a][\xi_j]^{[B]} \subset [\eta_j],$$

dann liegt eine Verbesserung der Meßintervalle vor, d.h. es ist nachgewiesen, daß durch die Mengendifferenz

$[\eta_j] \setminus [\alpha][\xi_j]^{[B]}$  keine Lösung einer korrespondierenden reellen Interpolationsaufgabe hindurchführt. Für die Praxis bedeutet dies, daß a posteriori Rückschlüsse auf die Genauigkeit der vorgenommenen Messungen gezogen werden können.

Wir wollen nun zeigen, daß die Gleichung (10) auch für allgemeinere - nicht notwendig lineare - monotone Funktionsstreifen gültig ist. Wir nehmen dazu an, daß Funktionsstreifen

$$[\varphi_k](a) = [\underline{\varphi}_k(a), \overline{\varphi}_k(a)] : D \rightarrow \mathbb{R}, \quad k=1(1)n,$$

wobei  $D$  eine zusammenhängende, nichtleere Teilmenge von  $\mathbb{R}$  ist,

gegeben sind. Es sei für  $i, j = 1(1)n$ ,  $i < j$ ,

$$F_{ij} := \{(a, B) \in D \times \mathbb{R} \mid ([\varphi_i](a) \cap [\varphi_j](a) \neq \emptyset) \\ \wedge (B \in [\varphi_m](a), m = i, j)\}.$$

Die Randfunktionen  $\underline{\varphi}_k$ ,  $\overline{\varphi}_k$  von  $[\varphi_k]$  seien stetig und mögen die folgenden beiden Eigenschaften besitzen:

(F 1) Zu  $i, j = 1(1)n$  mit  $i < j$  gibt es genau vier reelle Zahlen

$$\rho(\overline{I}, \underline{J}), \rho(\overline{I}, \overline{J}), \rho(\underline{I}, \underline{J}), \rho(\underline{I}, \overline{J}) \in D \text{ mit}$$

$$\overline{\varphi}_i(\rho(\overline{I}, \underline{J})) = \underline{\varphi}_j(\rho(\overline{I}, \underline{J})),$$

$$\overline{\varphi}_i(\rho(\overline{I}, \overline{J})) = \overline{\varphi}_j(\rho(\overline{I}, \overline{J})),$$

$$\underline{\varphi}_i(\rho(\underline{I}, \underline{J})) = \underline{\varphi}_j(\rho(\underline{I}, \underline{J})),$$

$$\underline{\varphi}_i(\rho(\underline{I}, \overline{J})) = \overline{\varphi}_j(\rho(\underline{I}, \overline{J})),$$

d.h. jede der beiden Funktionen  $\underline{\varphi}_i$  und  $\overline{\varphi}_i$  schneidet  
jede der beiden Funktionen  $\underline{\varphi}_j$  und  $\overline{\varphi}_j$  genau einmal.  
Darüberhinaus besitze  $\overline{\varphi}_i - \underline{\varphi}_j$  in  $\rho(\overline{i}, \underline{j})$  und  $\underline{\varphi}_i - \overline{\varphi}_j$   
in  $\rho(\underline{i}, \overline{j})$  einen strengen Vorzeichenwechsel.

(E 2) Für alle  $r \in \{\rho(\overline{i}, \underline{j}), \rho(\overline{i}, \overline{j}), \rho(\underline{i}, \underline{j}), \rho(\underline{i}, \overline{j})\}$  gilt:

$$\underline{\varphi}_i(r) < \overline{\varphi}_i(r) \text{ und } \underline{\varphi}_j(r) < \overline{\varphi}_j(r) .$$

Bemerkungen: 1.) Die Forderung (E 2) schließt aus, daß die  
Menge  $P_{ij}$  möglicherweise nicht beschränkt ist. Dies möge  
die folgende Skizze (vgl. Abb. 5) verdeutlichen. Es sei  
 $r := \rho(\overline{i}, \underline{j})$ ,  $\underline{\varphi}_j(r) = \overline{\varphi}_j(r)$  und  $\underline{\varphi}_i(x) < \underline{\varphi}_j(x) \leq \overline{\varphi}_j(x) < \overline{\varphi}_i(x)$   
für  $x < r$ .

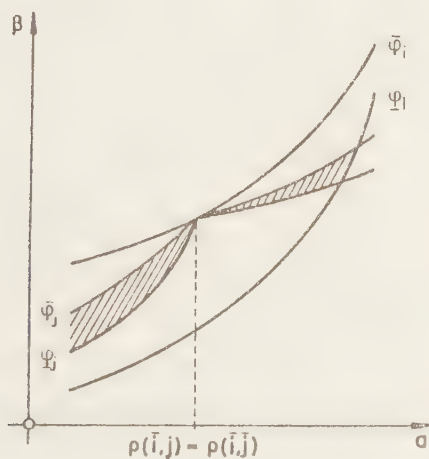


Abb. 5 Veranschaulichung zu  
Bemerkung 1

2) Nach Lemma 5.1 besitzen die Funktionen aus (11) die Eigenschaft (E 1). Es ist ferner nach (1.2e) für alle  $a \in \mathbb{R}$

$$\text{span}[\varphi_k](a) \geq \max\left\{\frac{1}{|\underline{x}_k|}, \frac{1}{|\bar{x}_k|}\right\} \text{span}[\eta_k],$$

(E 2) also erfüllt, wenn die Intervalle  $[\eta_i]$  sämtlich keine Punktintervalle sind.

Zur Abkürzung verzichten wir dort, wo keine Mißverständnisse zu befürchten sind, auf die Angabe der Indizes  $i$  und  $j$  und setzen für  $i < j$ :

$$r_1 := \rho(\bar{I}, \underline{J}), r_2 := \rho(\bar{I}, \bar{J}), r_3 := \rho(\underline{I}, \underline{J}), r_4 := \rho(\underline{I}, \bar{J}).$$

Lemma 5.2: Für  $i, j = 1(1)n$  mit  $i < j$  gilt:

$$\begin{aligned} \text{Aus} \quad & \forall a \in D : a < \rho(\bar{I}, \bar{J}) \Rightarrow \bar{\varphi}_i(a) \underset{(>)}{<} \bar{\varphi}_j(a) \\ \text{folgt} \quad & \forall a \in D : (\rho(\bar{I}, \bar{J}) < a \Rightarrow \bar{\varphi}_i(a) \underset{(<)}{>} \bar{\varphi}_j(a)) \\ & \wedge (a < \rho(\underline{I}, \underline{J}) \Rightarrow \underline{\varphi}_i(a) \underset{(>)}{<} \underline{\varphi}_j(a)). \end{aligned}$$

Beweis: Annahme: Es existiere ein  $a' \in D$  mit  $a' > r_2 = \rho(\bar{I}, \bar{J})$  und  $\bar{\varphi}_i(a') < \bar{\varphi}_j(a')$ , dann folgt wegen der Stetigkeit von  $\bar{\varphi}_i$  und  $\bar{\varphi}_j$  und (E 1)  $\bar{\varphi}_i(a) < \bar{\varphi}_j(a)$  für alle  $r_2 < a$  und daraus wegen (E 2)  $\underline{\varphi}_i(a) < \underline{\varphi}_j(a)$  für alle  $a \in D$  im Widerspruch zu (E 1).

Zum Beweis des zweiten Teils der Behauptung:

Angenommen, es gäbe ein  $a' \in D$  mit  $a' < r_3 = \rho(\underline{I}, \underline{J})$  und  $\underline{\varphi}_i(a') > \underline{\varphi}_j(a')$ , dann ist diese Ungleichung wieder für alle  $a < r_3$  erfüllt.

Wegen  $\underline{\varphi}_i(a) \leq \overline{\varphi}_i(a) < \overline{\varphi}_j(a)$  für alle  $a$ ,  $a < r_s$ ,

ist  $r_s \leq r_4 = \rho(\underline{i}, \overline{j})$ .

Gilt  $r_s = r_4$ , so ist

$$\underline{\varphi}_i(r_4) = \overline{\varphi}_j(r_4) = \overline{\varphi}_j(r_s) = \overline{\varphi}_i(r_s) = \overline{\varphi}_i(r_4)$$

im Widerspruch zu (E 1). Somit erhalten wir  $r_s < r_4$ .

Ferner ist wegen

$$\underline{\varphi}_j(a) < \underline{\varphi}_i(a) \leq \overline{\varphi}_i(a) \quad \text{für alle } a, a < r_s,$$

und nach (E 2)  $r_s < r_1 = \rho(\overline{i}, \underline{j})$ .

Mit dem ersten Teil der Behauptung erhält man

$$\underline{\varphi}_j(a) < \overline{\varphi}_i(a) \quad \text{für alle } a, r_s < a,$$

und somit  $r_1 < r_s$ .

Dies ergibt die Ungleichungskette  $r_s < r_1 < r_s < r_4$ .

Für alle  $a \in D$  mit  $r_s < a$  gilt

$$\underline{\varphi}_i(a) < \underline{\varphi}_j(a), \tag{13}$$

denn andernfalls wäre wieder  $\underline{\varphi}_j(a) < \overline{\varphi}_i(a)$  für alle  $a \in D$ .

Wegen  $r_s < r_4$  und  $\underline{\varphi}_j(r_4) < \overline{\varphi}_j(r_4) = \underline{\varphi}_i(r_4)$  ergibt sich ein Widerspruch zu (13).  $\blacktriangle$

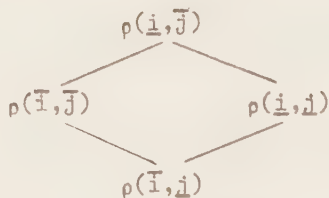
Im folgenden seien die Funktionen  $\underline{\varphi}_i$ ,  $\overline{\varphi}_i$ ,  $\underline{\varphi}_j$  und  $\overline{\varphi}_j$  so nummeriert, daß für  $i < j$  gilt

$$a < \rho(\overline{i}, \overline{j}) \Rightarrow \overline{\varphi}_i(a) < \overline{\varphi}_j(a). \tag{14}$$

Nach den Voraussetzungen über die Funktionen  $\underline{\varphi}_i$  und  $\overline{\varphi}_j$

genügt zum Nachweis von (14) schon die Angabe eines einzigen  $a \in D$  mit den erwünschten Eigenschaften. Im Falle, daß  $\bar{\varphi}_i$  isoton und  $\bar{\varphi}_j$  antiton sind, ist (14) stets erfüllt.

Lemma 5.3: a) Für die Abszissen der Schnittpunkte der Funktionen  $\varphi_i$ ,  $\bar{\varphi}_i$ ,  $\varphi_j$  und  $\bar{\varphi}_j$ ,  $i < j$ , gilt das Hasse-Diagramm 1)



(15)

b) Für alle  $i, j = 1(1)n$  mit  $i < j$  und alle  $a \in D$  gilt

$$\begin{aligned} \rho(\bar{i}, \underline{j}) \leq a &\Rightarrow \varphi_j(a) \leq \bar{\varphi}_i(a) \\ a \leq \rho(\underline{i}, \bar{j}) &\Rightarrow \bar{\varphi}_i(a) \leq \varphi_j(a) . \end{aligned}$$

(vergl. Abb. 6)

1)



bzw.



bedeuten  $x < y$ .

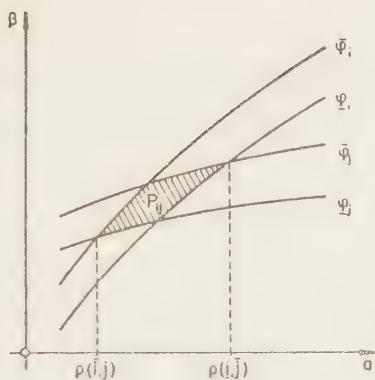


Abb. 6 Veranschaulichung  
zu Lemma 5.3

Beweis: zu a) Aus (14) folgt für alle  $a \in D$  mit  $a < r_2$  die Ungleichung  $\varphi_i(a) < \bar{\varphi}_j(a)$ , woraus  $r_2 < r_3$  folgt. Nach Lemma 5.2 gilt für alle  $a$  mit  $r_2 < a$  die Ungleichung  $\varphi_j(a) < \bar{\varphi}_i(a)$  und somit  $r_1 < r_3$ . Wir erhalten also  $r_1 < r_2 < r_3$ .

Entsprechend zeigt man die Ungleichungen  $r_1 < r_3 < r_4$ .

zu b) Es werde angenommen, es gäbe ein  $a' \in D$  mit  $r_1 = \rho(\bar{i}, j) \leq a'$  und  $\varphi_j(a') > \bar{\varphi}_i(a')$ ; dann folgt

$$\varphi_i(a) \leq \bar{\varphi}_i(a) < \varphi_j(a) \leq \bar{\varphi}_j(a) \text{ für alle } r_1 < a.$$

Damit ist aber  $r_4 = \rho(i, \bar{j}) < r_1$  im Widerspruch zu (15).

Den zweiten Teil der Behauptung zeigt man entsprechend.

Weitere Voraussetzung über die Funktionen  $\varphi_i$  und  $\bar{\varphi}_i$ .

Wir setzen nun für das folgende voraus, daß es eine Zerlegung der Indexmenge  $\{1, \dots, n\}$  in zwei disjunkte Teilmengen  $I$  und  $J$  gibt mit

$$\forall i \in I \forall j \in J : i < j$$

und

$$\{\varphi_i\} \text{ ist } \begin{cases} \text{isoton} & \text{für } i \in I, \\ \text{antiton} & \text{für } i \in J. \end{cases}$$

Möglicherweise ist entweder  $I$  oder  $J$  leer.

Lemma 5.4: Für die Ordinaten der Schnittpunkte der Funktionen  $\varphi_i$ ,  $\bar{\varphi}_i$ ,  $\varphi_j$  und  $\bar{\varphi}_j$ ,  $i < j$ , gelten die drei Hasse-Diagramme:

$i, j \in I$ :

$$\begin{array}{ccc}
 \varphi_i(\rho(\underline{i}, \bar{j})) & = & \bar{\varphi}_j(\rho(\underline{i}, \bar{j})) \\
 \varphi_i(\rho(\bar{i}, \bar{j})) = \bar{\varphi}_j(\rho(\bar{i}, \bar{j})) & & \varphi_i(\rho(\underline{i}, \underline{j})) = \varphi_j(\rho(\underline{i}, \underline{j})) \\
 \varphi_i(\rho(\bar{i}, \underline{j})) & = & \varphi_j(\rho(\bar{i}, \underline{j}))
 \end{array}$$

$i \in I, j \in J$ :

$$\begin{array}{ccc}
 \bar{\varphi}_i(\rho(\bar{i}, \bar{j})) & = & \bar{\varphi}_j(\rho(\bar{i}, \bar{j})) \\
 \bar{\varphi}_i(\rho(\bar{i}, \underline{j})) = \varphi_j(\rho(\bar{i}, \underline{j})) & & \varphi_i(\rho(\underline{i}, \bar{j})) = \bar{\varphi}_j(\rho(\underline{i}, \bar{j})) \\
 \varphi_i(\rho(\underline{i}, \underline{j})) & = & \varphi_j(\rho(\underline{i}, \underline{j}))
 \end{array}$$

$i, j \in J$ :

$$\begin{array}{ccc}
 \bar{\varphi}_i(\rho(\bar{i}, \underline{j})) & = & \varphi_j(\rho(\bar{i}, \underline{j})) \\
 \bar{\varphi}_i(\rho(\bar{i}, \bar{j})) = \bar{\varphi}_j(\rho(\bar{i}, \bar{j})) & & \varphi_i(\rho(\underline{i}, \underline{j})) = \varphi_j(\rho(\underline{i}, \underline{j})) \\
 \varphi_i(\rho(\underline{i}, \bar{j})) & = & \bar{\varphi}_j(\rho(\underline{i}, \bar{j}))
 \end{array}$$

(vgl. Abb.7).

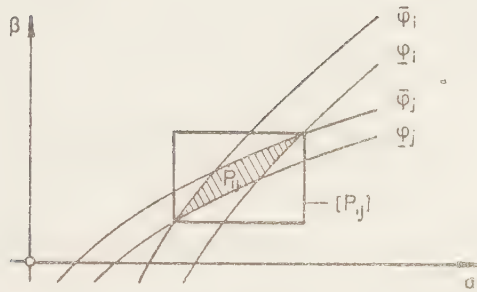


Abb. 7a Veranschaulichung zum Fall  $i, j \in I$

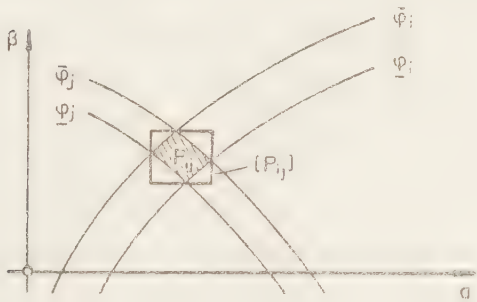


Abb. 7b Veranschaulichung zum Fall  $i \in I, j \in J$

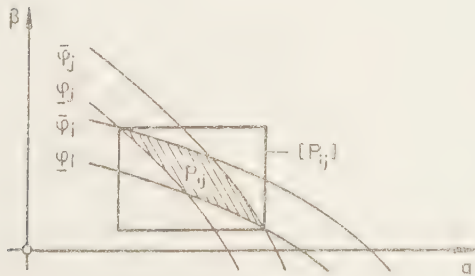


Abb. 7c Veranschaulichung zum Fall  $i, j \in J$

Beweis: Der Beweis für den Fall, daß  $i, j \in I$  oder  $i, j \in J$  sind, ergibt sich unmittelbar aus den Monotonieeigenschaften der Funktionen und (15);

zu  $i \in I$  und  $j \in J$  : Es gilt:

$$\begin{array}{c} \varphi_i(r_3) \leq \varphi_i(r_4) = \bar{\varphi}_j(r_4) \leq \bar{\varphi}_j(r_2) \\ \parallel \qquad \qquad \qquad \parallel \\ \varphi_j(r_3) \leq \varphi_j(r_1) = \bar{\varphi}_i(r_1) \leq \bar{\varphi}_i(r_2) . \end{array} \quad \blacktriangle$$

Definition ( $\underline{\rho}_{ij}$ ,  $\bar{\rho}_{ij}$ ,  $\underline{\rho}$ ,  $\bar{\rho}$ ,  $\underline{\beta}_{ij}$ ,  $\bar{\beta}_{ij}$ ,  $\underline{\beta}$ ,  $\bar{\beta}$ ):

Für  $i, j = 1(1)n$ ,  $i < j$ , definieren wir:

$$\underline{\rho}_{ij} := \rho(\bar{I}, \underline{j}) , \quad \bar{\rho}_{ij} := \rho(\underline{i}, \bar{J}) ,$$

$$\underline{\rho} := \sup \underline{\rho}_{ij} , \quad \bar{\rho} := \inf \bar{\rho}_{ij} ,$$

$$\underline{\beta}_{ij} := \begin{cases} \bar{\varphi}_i(\rho(\bar{I}, \underline{j})) & \text{für } i, j \in I, \\ \varphi_i(\rho(\underline{i}, \underline{j})) & \text{für } i \in I, j \in J, \\ \varphi_i(\rho(\underline{i}, \bar{j})) & \text{für } i, j \in J, \end{cases}$$

$$\bar{\beta}_{ij} := \begin{cases} \varphi_i(\rho(\underline{i}, \bar{j})) & \text{für } i, j \in I, \\ \bar{\varphi}_i(\rho(\bar{I}, \bar{j})) & \text{für } i \in I, j \in J, \\ \bar{\varphi}_i(\rho(\bar{I}, \underline{j})) & \text{für } i, j \in J, \end{cases}$$

$$\underline{\beta} := \sup \underline{\beta}_{ij} , \quad \bar{\beta} := \inf \bar{\beta}_{ij} .$$

Nach den Lemmata 5.3 und 5.4 gilt dann:

$$\underline{\rho}_{ij} \leq \bar{\rho}_{ij} \quad \text{und} \quad \underline{\beta}_{ij} \leq \bar{\beta}_{ij} .$$

Lemma 5.5: Für die optimale Intervalleinschließung  $\text{Intv } P_{ij}$  von  $P_{ij}$  gilt:

$$\text{Intv } P_{ij} = [(\underline{\rho}_{ij}, \underline{\beta}_{ij}), (\overline{\rho}_{ij}, \overline{\beta}_{ij})].$$

Der Beweis ergibt sich sofort aus der Monotonie der Randfunktionen und Lemma 5.4 . ▲

Wir definieren wie in (5)

$$P := \bigcap_{i < j} P_{ij} .$$

Die Gleichung (10) ist ein Sonderfall des folgenden Satzes:

Satz 5.1: Ist  $P \neq \emptyset$  , so ist  $\bigcap_{i < j} \text{Intv } P_{ij}$  die optimale Intervalleinschließung von  $P$ , d.h.

$$\text{Intv } P = \bigcap_{i < j} \text{Intv } P_{ij}$$

(vgl. Abb. 8).

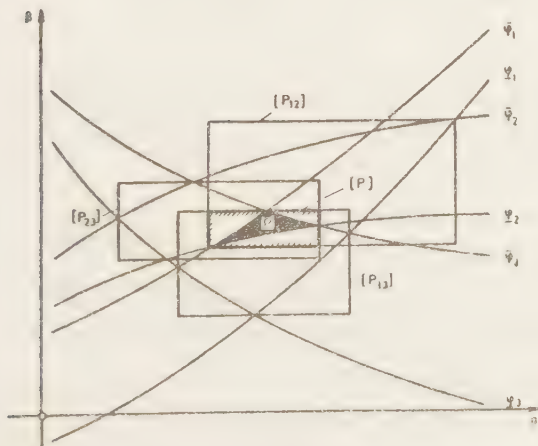


Abb. 8 Veranschaulichung des Sachverhaltes von Satz 5.1 im Falle  $n := 3$ ,  $I := \{1,2\}$ ,  $J := \{3\}$

Beweis: Nach Lemma 1.2 läßt Intv  $P_{ij}$  sich für  $i, j=1(1)n$ ,  
 $i < j$ , darstellen als

$$\text{Intv } P_{ij} = \text{Intv } \pi_1(P_{ij}) \times \text{Intv } \pi_2(P_{ij}).$$

Da  $\pi_1$  stetig und  $P_{ij}$  zusammenhängend und kompakt ist, folgt

$$\begin{aligned} \text{Intv } \pi_1(P_{ij}) &= \pi_1(P_{ij}) \\ &= \{a \in D \mid [\varphi_i](a) \cap [\varphi_j](a) \neq \emptyset\}. \end{aligned}$$

Es ist

$$\begin{aligned} P &= \{(a, B) \in D \times \mathbb{R} \mid ([\varphi_i](a) \cap [\varphi_j](a) \neq \emptyset, i, j=1(1)n) \\ &\quad \wedge \underline{\varphi}_k(a) \leq B \leq \overline{\varphi}_k(a), k=1(1)n\}, \end{aligned} \quad (16)$$

woraus folgt

$$\begin{aligned} \pi_1(\text{Intv } P) &\supseteq \pi_1(P) = \cap \pi_1(P_{ij}) = \cap \text{Intv } \pi_1(P_{ij}) \\ &= \cap \pi_1 \text{ Intv } (P_{ij}). \end{aligned}$$

Mit Lemma 1.3 folgt

$$\pi_1(\text{Intv } P) = \cap \pi_1 \text{ Intv } P_{ij}. \quad (17)$$

Im folgenden seien

$$\begin{aligned} R &:= \pi_1(\text{Intv } P) = [\underline{\rho}, \overline{\rho}], \\ B &= [\underline{b}, \overline{b}] := \pi_2(\text{Intv } P). \end{aligned}$$

Zum Beweis von  $\pi_2(\text{Intv } P) = \bigcap_{i < j} \pi_2(\text{Intv } P_{ij})$  betrachten wir nur den Fall, daß  $I \neq \emptyset$  und  $J \neq \emptyset$ . Ist entweder  $I$  oder  $J$  leer, so vereinfacht sich der nachstehende Beweis.

Es ist nach (16)  $\underline{b} = \inf_{a \in R} \sup_{i=1}^n \varphi_i(a)$ .

Für alle  $a \in R$  sei definiert:

$$\psi_1(a) := \sup_{i \in I} \varphi_i(a), \quad \psi_2(a) := \sup_{i \in J} \varphi_i(a),$$

dann sind  $\psi_1$  und  $\psi_2$  stetig auf  $R$ ,  $\psi_1$  ist als Supremum isotoner Funktionen isoton, und  $\psi_2$  als Supremum antitoner Funktionen antiton. Ferner sei definiert:  $\psi := \sup(\psi_1, \psi_2)$ , somit ist  $\underline{b} = \inf_{a \in R} \psi(a)$ .

Fall I)  $\psi_1(\hat{a}) = \psi_2(\hat{a})$  für ein  $\hat{a} \in R$ : Es gibt dann

$\lambda \in I, \mu \in J$  mit

$$\psi_1(\hat{a}) = \varphi_\lambda(\hat{a}) = \varphi_\mu(\hat{a}) = \psi_2(\hat{a});$$

für alle  $a \in R$  gilt:

$$\text{für } a \leq \hat{a}: \psi_1(a) \leq \psi_1(\hat{a}) = \psi_2(\hat{a}) \leq \psi_2(a) = \psi(a),$$

$$\text{für } \hat{a} \leq a: \psi_2(a) \leq \psi_2(\hat{a}) = \psi_1(\hat{a}) \leq \psi_1(a) = \psi(a),$$

also folgt

$$\underline{b} = \inf_{a \in R} \psi(a) = \psi(\hat{a}) = \varphi_\mu(\varphi_\lambda(\underline{\lambda}, \underline{\mu})) = \underline{B}_{\lambda\mu}.$$

Fall II)  $\psi_1(a) \neq \psi_2(a)$  für alle  $a \in R$ : Es ist  $\psi$  isoton oder antiton.

Fall a)  $\psi = \psi_1$ , also ist  $\psi$  isoton und es gilt

$$\inf \psi(a) = \psi(\underline{\rho}); \text{ es gibt ein } \lambda \in I \text{ mit } \psi(\underline{\rho}) = \varphi_\lambda(\underline{\rho})$$

und  $\mu, \nu$  mit  $\mu < \nu$  und  $\underline{\rho} = \underline{\rho}_{\mu\nu}$ , also ist

$$\varphi_\mu(\underline{\rho}) = \varphi_\nu(\underline{\rho}) \leq \varphi_\lambda(\underline{\rho}), \quad (18)$$

insbesondere ist wegen (E 2)  $\mu \neq \lambda$ .

1)  $\underline{\mu} < \underline{\lambda}$ :

Es ist also  $\bar{\varphi}_{\underline{\mu}}$  isoton. Wegen  $\underline{\rho} = \underline{\rho}_{\underline{\mu}\underline{\nu}}$  ist  $\underline{\rho}_{\underline{\mu}\underline{\lambda}} \leq \underline{\rho}_{\underline{\mu}\underline{\nu}}$ , woraus mit Lemma 5.3b)

$$\varphi_{\underline{\lambda}}(\underline{\rho}_{\underline{\mu}\underline{\nu}}) \leq \bar{\varphi}_{\underline{\mu}}(\underline{\rho}_{\underline{\mu}\underline{\nu}})$$

folgt; zusammen mit (18) ergibt sich

$$\bar{\varphi}_{\underline{\mu}}(\underline{\rho}) = \varphi_{\underline{\lambda}}(\underline{\rho})$$

und damit  $\underline{b} = \inf_{a \in R} \psi(a) = \varphi_{\underline{\lambda}}(\rho(\bar{\underline{\mu}}, \underline{\lambda}))$ ,

$$\text{also } \underline{b} = \underline{\beta}_{\underline{\mu}\underline{\lambda}}.$$

2)  $\underline{\lambda} < \underline{\mu}$ :

Es ist  $\underline{\rho}_{\underline{\mu}\underline{\nu}} \leq \bar{\rho}_{\underline{\lambda}\underline{\mu}}$ , denn sonst wäre P leer; mit Lemma 5.3b) folgt daraus

$$\varphi_{\underline{\lambda}}(\underline{\rho}_{\underline{\mu}\underline{\nu}}) \leq \bar{\varphi}_{\underline{\mu}}(\underline{\rho}_{\underline{\mu}\underline{\nu}}),$$

zusammen mit (18) ergibt sich wieder

$$\varphi_{\underline{\lambda}}(\underline{\rho}) = \bar{\varphi}_{\underline{\mu}}(\underline{\rho});$$

(a)  $\underline{v} \in I$ :

es ist  $\varphi_{\underline{v}}$  - und wegen  $\underline{\mu} < \underline{v}$  auch  $\bar{\varphi}_{\underline{\mu}}$  - isoton; analog Fall 1) ergibt sich  $\underline{b} = \underline{\beta}_{\underline{\mu}\underline{v}}$ ;

(B)  $\underline{v} \in J$ :

es ist  $\varphi_{\underline{v}}$  antiton und  $\underline{b} = \varphi_{\underline{\lambda}}(\rho(\underline{\lambda}, \underline{v}))$ ,  
also  $\underline{b} = \underline{\beta}_{\underline{\lambda}\underline{v}}$ .

In allen Fällen von II) a) ist also

$$\underline{b} \leq \underline{\beta} = \sup \underline{\beta}_{ij}.. \quad (19)$$

Fall b)  $\psi = \psi_2$ , also ist  $\psi$  antiton: analog zu Fall a) zeigt man, daß (19) gilt.

Entsprechend erhält man  $\bar{B} \leq \bar{b} = \sup_{a \in R} \inf_{i=1}^n \bar{\varphi}_i(a)$ , es gilt also

$$\bigcap_{i < j} \pi_2(\text{Intv } P_{ij}) = [\underline{\beta} \quad \bar{\beta}] \subseteq \pi_2(\text{Intv } P) .$$

Hieraus ergibt sich mit Lemma 1.3 und (17) die behauptete Gleichung des Satzes.  $\blacktriangle$

Der folgende Satz gibt ein Kriterium an, mit Hilfe dessen schon während der Berechnung des einschließenden Intervalls für den Parameter  $a$  entschieden werden kann, ob  $P = \emptyset$  ist.

Satz 5.2: Es gilt

$P = \emptyset$  genau dann, wenn es  $1 \leq i, j, k, m \leq n$   
mit  $\bar{\rho}_{km} < \underline{\rho}_{ij}$  gibt.

Beweis: Es ist nur die Richtung " $\Rightarrow$ " zu zeigen.

Es gelte  $\underline{\rho}_{ij} \leq \bar{\rho}_{km}$  für  $i, j, k, m = 1(1)n$ ,  $i < j$  und  $k < m$ ; (20)  
zu zeigen ist, daß  $P \neq \emptyset$  ist.

Nach (20) ist  $\underline{\rho} = \sup \underline{\rho}_{ij} \leq \inf \bar{\rho}_{ij} = \bar{\rho}$ , woraus nach Lemma 5.3 b) für alle  $a \in [\underline{\rho}, \bar{\rho}]$  und  $i, j = 1(1)n$ ,  $i < j$ ,

$$\underline{\varphi}_j(a) \leq \bar{\varphi}_i(a) \wedge \underline{\Psi}_i(a) \leq \bar{\varphi}_j(a) ,$$

also  $[\varphi_i](a) \cap [\varphi_j](a) \neq \emptyset$  folgt, damit ist aber  $P \neq \emptyset$ .  $\blacktriangle$

Wird bei der Bestimmung von  $\text{Intv } P$   $\underline{\rho}$  bzw.  $\bar{\rho}$  vor  $\underline{\beta}$  bzw.  $\bar{\beta}$  berechnet, dann reduziert sich möglicherweise der Aufwand zur Bestimmung von  $\underline{\beta}$  bzw.  $\bar{\beta}$  auf die Berechnung eines einzigen  $\underline{\beta}_{ij}$  bzw.  $\bar{\beta}_{ij}$  :

Satz 5.3: Es gilt, falls  $P \neq \emptyset$  ist:

$$\forall i, j \in I: (\underline{p}_{ij} = \underline{p} \Rightarrow \underline{B}_{ij} = \underline{B}) \wedge (\overline{p}_{ij} = \overline{p} \Rightarrow \overline{B}_{ij} = \overline{B}),$$

$$\forall i, j \in J: (\overline{p}_{ij} = \overline{p} \Rightarrow \underline{B}_{ij} = \underline{B}) \wedge (\underline{p}_{ij} = \underline{p} \Rightarrow \overline{B}_{ij} = \overline{B}).$$

Bemerkung: 1) Eine entsprechende Aussage für den Fall  $i \in I$  und  $j \in J$  ist i.a. falsch.

2) Die Voraussetzung  $P \neq \emptyset$  kann nach der Berechnung von  $\underline{p}$  und  $\overline{p}$  wegen des Kriteriums aus Satz 5.2 verifiziert werden.

Beweis: Wir wollen den Beweis nur für den ersten Teil der ersten Behauptung führen, den anderen Teil und die zweite Behauptung zeigt man entsprechend.

Es seien  $\mu, \nu \in I$ , also  $\overline{\varphi}_\mu$  und  $\varphi_\nu$  isoton, und  $\underline{p}_{\mu\nu} = \underline{p}$ . Wir verwenden die Bezeichnungen und Ergebnisse des Beweises zu Satz 5.1 und treffen die gleiche Fallunterscheidung wie dort.

Fall I)  $\psi_1(\hat{a}) = \psi_2(\hat{a})$  für ein  $\hat{a} \in R$ :

Es gibt ein  $\lambda \in J$  mit

$$\underline{\varphi}_\lambda(\hat{a}) = \inf_{a \in R} \psi(a) = \underline{B}. \quad (20)$$

Aus  $\underline{p}_{\mu\nu} = \underline{p}$  folgt  $\underline{p}_{\mu\lambda} \leq \underline{p}_{\mu\nu}$  und damit für alle  $a \in D$  mit  $\underline{p}_{\mu\nu} \leq a$ :

$$\begin{aligned} \underline{\varphi}_\lambda(a) &\leq \underline{\varphi}_\lambda(\underline{p}_{\mu\lambda}) = \overline{\varphi}_\mu(\underline{p}_{\mu\lambda}) \leq \overline{\varphi}_\mu(\underline{p}_{\mu\nu}) = \varphi_\nu(\underline{p}_{\mu\nu}) \\ &\leq \varphi_\nu(a) \leq \psi_1(a). \end{aligned} \quad (22)$$

Insbesondere folgt für  $a := \hat{a}$  mit (21)

$$\underline{\varphi}_\lambda(\hat{a}) \leq \underline{\varphi}_\nu(\hat{a}) \leq \psi_1(\hat{a}) = \psi_2(\hat{a}) = \underline{\varphi}_\lambda(\hat{a})$$

damit ist für alle  $a \in D$ ,  $\underline{\rho}_{\mu\nu} \leq a \leq \hat{a}$ ,

$$\underline{\varphi}_{\nu}(\hat{a}) = \underline{\varphi}_{\lambda}(\hat{a}) \leq \underline{\varphi}_{\lambda}(a) \leq \underline{\varphi}_{\nu}(a) \leq \underline{\varphi}_{\nu}(\hat{a}), \quad (22)$$

also  $\underline{\varphi}_{\nu}(a) = \underline{\varphi}_{\lambda}(a)$  für alle  $a$ ,  $\underline{\rho}_{\mu\nu} \leq a \leq \hat{a}$ ,

und wegen (E 1)  $\hat{a} = \underline{\rho}_{\mu\nu} = \underline{\rho}$ .

Damit ist

$$\underline{\beta} = \underline{\varphi}_{\lambda}(\hat{a}) = \underline{\varphi}_{\lambda}(\underline{\rho}) = \underline{\varphi}_{\nu}(\underline{\rho}) = \underline{\varphi}_{\nu}(\underline{\rho}_{\mu\nu}) = \underline{\beta}_{\mu\nu}.$$

Fall II)  $\psi_1(a) \neq \psi_2(a)$  für alle  $a \in R$ :

Fall a)  $\psi = \psi_1$ , also gilt  $\underline{\beta} = \psi(\underline{\rho})$  und mit den Ergebnissen des Beweises zu Satz 5.1 ist

$$\underline{\beta}_{\mu\nu} = \overline{\varphi}_{\mu}(\underline{\rho}) = \underline{\varphi}_{\nu}(\underline{\rho}) = \psi(\underline{\rho}) = \underline{\beta}.$$

Fall b)  $\psi = \psi_2$ , also ist  $\underline{\beta} = \psi(\overline{\rho})$ ;

es sei  $\lambda \in J$  mit  $\underline{\varphi}_{\lambda}(\overline{\rho}) = \psi(\overline{\rho})$ . Wie in Fall I) gilt für alle  $a \in D$  mit  $\underline{\rho}_{\mu\nu} \leq a$

$$\underline{\varphi}_{\lambda}(a) \leq \underline{\varphi}_{\nu}(a) \leq \psi_1(a),$$

es ist also  $\psi(\overline{\rho}) = \psi_2(\overline{\rho}) = \underline{\varphi}_{\lambda}(\overline{\rho}) \leq \psi_1(\overline{\rho})$ ,

also  $\psi_1(\overline{\rho}) = \psi_2(\overline{\rho})$  im Widerspruch zur Voraussetzung

$\psi_1(a) \neq \psi_2(a)$  für alle  $a \in R$ . ▲

Eine Umkehrung des Satzes 5.3, d.h. eine mögliche Verminderung des Aufwandes bei der Berechnung von  $\underline{\rho}$  bzw.  $\overline{\rho}$ , wenn  $\underline{\beta}$  bzw.  $\overline{\beta}$  zuerst berechnet wird, liefert der folgende Satz:

Satz 5.4: Falls  $P \neq \emptyset$  ist, gilt:

$$\begin{aligned} \forall i, j \in I: & ((\underline{\varphi}_j \text{ streng isoton} \wedge \underline{\beta}_{ij} = \underline{\beta}) \Rightarrow \underline{\rho}_{ij} = \underline{\rho}) \wedge \\ & (\overline{\varphi}_j \text{ streng isoton} \wedge \overline{\beta}_{ij} = \overline{\beta}) \Rightarrow \overline{\rho}_{ij} = \overline{\rho}), \\ \forall i, j \in J: & ((\underline{\varphi}_i \text{ streng antiton} \wedge \underline{\beta}_{ij} = \underline{\beta}) \Rightarrow \underline{\rho}_{ij} = \underline{\rho}) \wedge \\ & (\overline{\varphi}_i \text{ streng antiton} \wedge \overline{\beta}_{ij} = \overline{\beta}) \Rightarrow \underline{\rho}_{ij} = \underline{\rho}). \end{aligned}$$

Der Beweis verläuft ähnlich wie der zum Satz 5.3 und soll hier unterdrückt werden.

#### Numerisches Beispiel

Wir betrachten die Intervallinterpolation vom Typ I mit den Funktionen  $\alpha \xi^B$ .

Es sei  $n = 5$  und

$$\begin{aligned} \xi_1 &:= 0.25, & [\eta_1] &:= [1.25, 0.25], \\ \xi_2 &:= 0.5, & [\eta_2] &:= [0.25, 0.5], \\ \xi_3 &:= 1.5, & [\eta_3] &:= [0.5, 2], \\ \xi_4 &:= 2, & [\eta_4] &:= [1, 4], \\ \xi_5 &:= 3, & [\eta_5] &:= [4, 10]. \end{aligned}$$

Die optimale Intervalleinschließung der Menge  $\mathcal{M}$  aller Parameterpaare  $(\sigma, \beta) \in \mathbb{R}_+ \times \mathbb{R}$  mit  $\alpha \xi_k^B \in [\eta_k]$ ,  $k=1(1)n$ , läßt sich unter Verwendung von Satz 5.1 gemäß (6)-(9) berechnen. Zu betrachten ist, daß in (6) die Gleichheit gilt.

Das im Anhang aufgeführte Triplex-ALGOL-Programm liefert für die Intervalle  $[a]$  und  $[B]$  aus (12) <sup>1)</sup>

$$[a] = [0.863210, 1.211844] ,$$

$$[B] = [1.158367, 1.555829] .$$

Die Intervalle  $[\eta_k]$  konnten zu den folgenden Intervallen  $[\eta'_k]$  verbessert werden

$$[\eta'_1] := [0.125 , 0.243818] ,$$

$$[\eta'_2] := [0.293447, 0.5 ] ,$$

$$[\eta'_3] := [1.378640, 2 ] ,$$

$$[\eta'_4] := [1.921755, 3.566288] ,$$

$$[\eta'_5] := [4 , 6.708405] .$$

Die Spannen der Intervalle  $[a]$ ,  $[B]$ ,  $[\eta'_k]$  sind zwar etwas größer als die Spannen der entsprechenden Intervalle aus [21]. Dies wird verursacht durch maschineninterne Darstellungsfehler. Das im Anhang aufgeführte Programm ist jedoch auch zur Intervallinterpolation vom Typ II einsetzbar und in seinem Aufbau übersichtlicher und kürzer als das Programm aus [21].

---

1) Die Rechnungen wurden auf der UNIVAC 1100/81 (mit 8 Ziffern) des Rechenzentrums der Universität Freiburg i.Br. durchgeführt.

## 5.2 Intervallinterpolation mit den Funktionen

$$\underline{\alpha_1 \exp(\beta_1 \xi) + \alpha_2 \exp(\beta_2 \xi)}$$

In Verallgemeinerung der Problemstellung des vorangegangenen Abschnitts betrachten wir noch kurz die Interpolation mit Exponentialsummen

$$\varphi(\xi) := \sum_{i=1}^m \alpha_i \exp(\beta_i \xi) , \quad \xi \in \mathbb{R} .$$

Es seien die äquidistanten Stützstellen

$$\xi_k := \xi_0 + kh , \quad k=1(1)2m , \quad h > 0 ,$$

vorgegeben. O.B.d.A. sei  $\xi_0 := -h$  gewählt; es ist also  $\xi_1 = 0$  .

Wir betrachten zunächst das reelle Interpolationsproblem:<sup>1)</sup> Gesucht sind die Koeffizienten  $\alpha_i$  und die Exponenten  $\beta_i$  , so daß gilt

$$\varphi(\xi_k; \alpha_1, \dots, \alpha_m, \beta_1, \dots, \beta_m) = \eta_k , \quad k=1(1)2m .$$

Die Auflösung eines nichtlinearen Gleichungssystems zur Bestimmung der  $\alpha_i$  und  $\beta_i$  vermeidet das Verfahren von Prony [41] (s. etwa auch [32]), das sich unmittelbar aus der folgenden Beobachtung ergibt:

1) Wir beschränken uns hier auf die reelle Interpolation. Eine Erweiterung auf den komplexen Fall ist möglich.

Besitzt das lineare Gleichungssystem

$$\begin{pmatrix} \eta_1 & \eta_2 & \dots & \eta_m \\ \eta_2 & \eta_3 & \dots & \eta_{m+1} \\ \vdots & \vdots & & \vdots \\ \eta_m & \eta_{m+1} & \dots & \eta_{2m-1} \end{pmatrix} \cdot \begin{pmatrix} \zeta_1 \\ \zeta_2 \\ \vdots \\ \zeta_m \end{pmatrix} = \begin{pmatrix} -\eta_{m+1} \\ -\eta_{m+2} \\ \vdots \\ -\eta_{2m} \end{pmatrix} \quad (23)$$

die eindeutig bestimmte Lösung  $(\zeta_1, \dots, \zeta_m)^T$  und das Polynom

$$q(u) := u^m + \zeta_m u^{m-1} + \dots + \zeta_2 u + \zeta_1 \quad (24)$$

zu verschiedenen reellen Nullstellen  $u_1, \dots, u_m \in \mathbb{R}$  und gilt für die Exponenten  $B_i$

$$B_i = (\ln u_i)/h, \quad i=1(1)m. \quad (25)$$

Zur Bestimmung der Koeffizienten  $\zeta_j$  läßt sich ein lineares Gleichungssystem aufstellen.

Sind die Matrizen

$$A_{ij} = \frac{1}{h} \ln u_i, \quad i, j=1, \dots, m, \quad P = (p_{ij})$$

symmetrisch definit, dann ist das Verfahren durchführbar. Insbesondere sind alle  $B_i$  reell und paarweise verschieden (vgl. [29], S. 75).

Sei  $II$  das vom Interpolationsproblem der Typ I-1) abgeleitete ein (22) erfüllendes lineares Intervallinterpolationsproblem

- 1) Das Problem vom Typ II ist wegen der Äquidistanz der Stützstellen mit dem Verfahren von Prony nicht lösbar.

(s. Abschnitt 4.1) zu lösen. Mit den Komponenten einer Lösungseinschließung erhält man dann die Koeffizienten  $[c_i]$  von

$$u^m - \sum_{j=1}^m [c_j] u^{j-1} : R \rightarrow II(R) .$$

Nun sind noch die Nullstellen  $u_i(\{\xi_j\})$  sämtlicher Polynome

$$u^m - \sum_{j=1}^m c_j u^{j-1} \quad \text{mit } c_j \in [c_j], \quad j=1(1)m ,$$

in paarweise disjunkte Intervalle  $[u_i]$  einzuschließen

$$u_i(\{\xi_j\}) \in [u_i] , \quad i=1(1)n ,$$

um Schranken für die Exponenten zu erhalten. Dies ist grundsätzlich beispielsweise mit dem Verfahren von Bachmann [3] möglich. Wegen der bekannten Sensitivität der Exponenten gegenüber kleinen Störungen ("exponential approximation is a notoriously delicate enterprise", [6], S. 242) muß jedoch mit erheblich großen Spannen der Intervalle  $[u_i]$  gerechnet werden. Die für das reelle Problem in [46], S. 43-48, und [26], Kap.7, vorgeschlagene Ermittlung der Exponenten  $B_i$  mit dem QD-Algorithmus von Rutishauser ist nicht ohne weiteres auf das Intervall-Problem übertragbar, da u.a. wie Beispiele zeigen, der QD-Algorithmus bei intervallmäßiger Durchführung abbrechen kann.

Im Fall  $m = 2$ ,

$$p(\xi) = \alpha_1 \exp B_1 \xi + \alpha_2 \exp B_2 \xi ,$$

können die Nullstellen von  $q$  in (24) explizit berechnet werden. Wir geben nachstehend einen Algorithmus für die Intervallinterpolation für  $m = 2$  an. Es seien  $n \geq 4$  reelle Intervalle  $[\eta_k]$  gegeben und Intervalle  $[\alpha_i], [\beta_i]$ ,  $i = 1, 2$ , gesucht, die die Koeffizienten und Exponenten jeder Funktion  $\varphi$  aus (26) mit  $\varphi(\xi_k) \in [\eta_k]$ ,  $k=1(1)n$  enthalten, falls derartige Funktionen existieren.

Der Algorithmus folgt dem Verfahren von Prony für die reelle Interpolation.

Im folgenden bedeutet "Abbruch", daß der Algorithmus abubrechen ist. In diesem Falle kann nicht entschieden werden, ob es eine reelle Funktion  $\varphi$  der Gestalt (26) gibt mit  $\varphi(\xi_k) \in [\eta_k]$ ,  $k=1(1)n$ . Bricht der Algorithmus nicht frühzeitig, d.h. vor Erreichen von (e) ab, dann liefert er Schranken für die Parameter jeder derartigen Interpolierenden.

#### Algorithmus für die Intervallinterpolation vom Typ I

mit  $\alpha_1 \exp B_1 \xi + \alpha_2 \exp B_2 \xi$

$$(a) \quad [\delta_k] := [\eta_k][\eta_{k+2}] - [\eta_{k+1}]^2 \quad \text{für } k=1(1)n-2$$

$$\text{falls } 0 \in [\delta_j] \quad \text{für ein } j \in \{1, \dots, n-3\} \rightarrow \text{Abbruch.} \quad (28)$$

$$(b) \quad \left. \begin{aligned} [\zeta_k^{(1)}] &:= [\delta_{k+1}] / [\delta_k] \\ [\zeta_k^{(2)}] &:= ([\eta_{k+1}][\eta_{k+2}] - [\eta_k][\eta_{k+3}]) / [\delta_k] \end{aligned} \right\} \quad \text{für } k=1(1)n-3$$

$$[\zeta^{(p)}] := \bigcap_{k=1}^{n-3} [\zeta_k^{(p)}] \quad \text{für } p = 1, 2,$$

falls  $[\zeta^{(p)}] = \emptyset$  für  $p = 1$  oder  $p = 2 \rightarrow$  Abbruch.

$$(c) \quad [r] := [\underline{r}, \bar{r}] := \frac{1}{4} [\zeta^{(2)}]^2 - [\zeta^{(1)}] ,$$

falls  $\text{sign}[r] < 1$

→ Abbruch.

$$[u_{1/2}] := -\frac{1}{2} [\zeta^{(2)}] \pm [V\underline{r}, V\bar{r}] ,$$

falls  $[u_p] = [\underline{u}_p, \bar{u}_p] \leq 0$  für  $p=1$  oder  $p=2$  → Abbruch.

$$[u_p] := [\sup\{0, \underline{u}_p\}, \bar{u}_p] \quad \text{für } p=1,2$$

(29)

falls  $[u_1] \cap [u_2] \neq \emptyset$

→ Abbruch.

$$(d) \quad [\gamma] := [u_2] - [u_1] ,$$

falls  $0 \in [\gamma]$

→ Abbruch

$$[\alpha_1] := ([n_1][u_2] - [n_2])/[\gamma] ,$$

$$[\alpha_2] := ([n_2] - [n_1][u_1])/[\gamma] ,$$

$$(e) \quad [B_p] := \ln[u_p]/h , \quad p = 1,2 .$$

Bemerkungen: 1.) Wird in (29)  $\underline{u}_p := 0$  gesetzt, so ergibt sich in Schritt (e)  $B_p := -\infty$ , man erhält also ein unbeschränktes reelles Intervall.

2.) Bricht der Algorithmus bei (28) ab, dann ist zu prüfen, ob die zu interpolierende Funktion nicht die Gestalt

$$\psi(\xi) := \alpha \exp B \xi$$

besitzt. Gilt nämlich in diesem Falle für ein  $k \in \{1, \dots, n-2\}$

$$\psi(\xi_{k+j}) \in [\eta_{k+j}], \quad j=0,1,2 ,$$

dann ist

$$\psi(\xi_k) \psi(\xi_{k+2}) - \psi(\xi_{k+1})^2 \in [\delta_k]$$

und damit

$$\alpha^2 \exp(\beta k h) \exp(\beta(k+2)h) - \alpha^2 \exp(\beta(2k+2)h) = 0 \in [\delta_k]$$

und der Algorithmus bricht bei (28) ab.

3.) Es läßt sich gegebenenfalls wieder eine Verbesserung der Meßintervalle erzielen, falls man noch

$$[\alpha_1][u_1]^k + [\alpha_2][u_2]^k, \quad k=1(1)n,$$

berechnet.

### Numerische Beispiele

Der vorstehende Algorithmus wurde in Triplex-ALGOL auf der UNIVAC 1100/81 des Rechenzentrums der Universität Freiburg i.Br. implementiert und lieferte die folgenden Ergebnisse:

1.) für  $n = 10$ ,  $h = 1$  und

$$[\eta_k] := \exp(\xi_k) + \exp(-\xi_k) + \delta[-1,1]$$

| $\delta$  | $[\alpha_1]$              | $[\alpha_2]$              | $[\beta_1]$               | $[\beta_2]$                  |
|-----------|---------------------------|---------------------------|---------------------------|------------------------------|
| $10^{-7}$ | [0.9999940,<br>1.0000061] | [0.9999938,<br>1.0000062] | [0.9999987,<br>1.0000014] | [-1.0000096,<br>- 0.9999907] |
| $10^{-5}$ | [0.9996278,<br>1.0003729] | [0.9996172,<br>1.0003823] | [0.9999208,<br>1.0000793] | [-1.0005840,<br>-0.9994165]  |

Tab. 7 Beispiel 1 zur Exponentialsummeninterpolation

2.) für  $n = 10$ ,  $h = 0.1$  und

$$[\eta_k] := 2\exp(2\xi_k) - 2\exp(10\xi_k) + \delta[-1,1]$$

| $\delta$  | $[\alpha_1]$              | $[\alpha_2]$                | $[\beta_1]$               | $[\beta_2]$                      |
|-----------|---------------------------|-----------------------------|---------------------------|----------------------------------|
| $10^{-7}$ | [1.9999870,<br>2.0000142] | [-2.0000139,<br>-1.9999871] | [1.9999611,<br>2.0000421] | [0.9999980,<br>1.0000019]<br>*10 |
| $10^{-5}$ | [1.9991763,<br>2.0008359] | [-2.0008259,<br>-1.9991854] | [1.9975461,<br>2.0024825] | [0.9998879,<br>1.0001105]<br>*10 |

Tab. 8 Beispiel 2 zur Exponentialsummeninterpolation

Diese Schranken sind angesichts der schon erwähnten Störanfälligkeit der Exponentialsummeninterpolation recht eng.

## Anhang: Triplex-ALGOL 60 - Programme

Hinsichtlich der Intervallprogrammiersprache Triplex-ALGOL 60 sei auf [55] verwiesen. Zusätzlich zu den in Triplex-ALGOL 60 vorgesehenen Standardfunktionen sind bei dem Triplex-ALGOL-Compiler des Rechenzentrums der Universität Freiburg i.Br. u.a. die folgenden Standardfunktionen verfügbar (X und Y seien formale Parameter vom Typ triplex):

| Funktion       | Resultattyp    | Resultat                                                                    |
|----------------|----------------|-----------------------------------------------------------------------------|
| SPAN (X)       | <u>real</u>    | sup(X)- inf(X)                                                              |
| DISJUNCT (X,Y) | <u>boolean</u> | <u>true</u> ,falls $X \cap Y = \emptyset$ gilt,<br><u>false</u> sonst       |
| INTSCT (X,Y)   | <u>triplex</u> | $X \cap Y$ , <u>main</u> -Wert ist die<br>Mitte des Resultat-<br>intervalls |

### 1. Die Prozedur IFFT

Die Prozedur IFFT dient zur intervallmäßigen Durchführung der FFT und zur Bestimmung von Einschließungsintervallen für die Koeffizienten des trigonometrischen Interpolationspolynoms unter Verwendung der Kreisscheibenarithmetik. Es werden die folgenden beiden Triplex-Funktionsprozeduren POINT und PSP benötigt.

POINT: bildet aus einer Maschinenzahl das einpunktige Maschinenintervall, welches nur aus dieser Zahl besteht.

PSP: bildet für ein Intervall  $X$  das einpunktige Maschinenintervall, bestehend aus  $\text{span } X$  ( $\hat{=}$  POINT (SPAN( $X$ ))).

Die Prozedur IFFT benötigt als Eingabeparameter

| Name     | Typ                  |
|----------|----------------------|
| $N, P$   | <u>integer</u>       |
| SCHAL    | <u>boolean</u>       |
| $YR, YI$ | <u>triplex array</u> |

Der Vektor

$$(YR(J) + iYI(J))_{J=0}^{N-1} = ([\eta_{2J}] + i[\eta_{2J+1}])_{J=0}^{N-1}, N=2^P, P \geq 3,$$

wird zunächst komponentenweise in Kreisscheiben gemäß (3.21) eingeschlossen. Der so gebildete Kreisscheibenvektor wird 'schnell' transformiert und der Ergebnisvektor komponentenweise gemäß (3.22) in Rechtecke eingeschlossen.  $YR(J)$  bzw.  $YI(J)$  wird dann mit dem Real- bzw. Imaginärteil der  $J$ -ten Komponente des bitinvertierten Rechteckvektors überschrieben ( $J=0(1)N-1$ ).

Im Fall  $\text{SCHAL} := \text{true}$  werden bei Beendigung des Prozeduraufrufes die Intervalle  $YR(J)$ ,  $YI(J)$ ,  $J=0(1)N$ , übergeben. Zu beachten ist, daß die Division durch  $N$  nicht vorgenommen wurde und daß die Komponenten  $YR(N)$  und  $YI(N)$  ohne Bedeutung sind.

Im Fall `SCHAL := false` sind nach Beendigung des Prozeduraufrufs die Komponenten `YR(J)` und `YI(J)` mit den Einschließungsintervallen  $[a_J]$  und  $[B_J]$  ( $J=0(1)N$ ) für die Koeffizienten des trigonometrischen Interpolationspolynoms überschrieben.

Der Aufbau der FFT folgt dem des Programms aus [49]. Die Transformation gemäß (24) und (25) bzw. (26) erfolgt im wesentlichen in Anlehnung an [47], S. 437.

```
•TRIPLEX••PROCEDURE•POINT(X);
  •VALUE•X;•REAL•X;
  POINT:=COMPOSE(X,X,X);
```

```
•TRIPLEX••PROCEDURE•PSP(X);
  •VALUE•X;•TRIPLEX•X;
  PSP:=POINT(SPAN(X));
```

```
•PROCEDURE•IFFT(N,P,SCHAL,YR,YI);
```

```
  •VALUE•N,P,SCHAL;•BOOLEAN•SCHAL;
```

```
  •COMMENT• FALLS NUR DIE FFT DURCHGEFUEHRT WERDEN SOLL, IST SCHAL
    = TRUE ZU SETZEN. ANDERNFALLS WERDEN INTERVALLE BERECHNET,
    WELCHE DIE KOEFFIZIENTEN DES TRIGONOMETRISCHEN INTERPOLATIONS-
    POLYNOMS ENTHALTEN;
```

```
  •INTEGER•N,P;•TRIPLEX••ARRAY•YR,YI;
  •BEGIN;••INTEGER•NI,NH,I,J,K,INDDIF,NGR,IG,INDA,IGR,IP,JH,KP,I1,I2;
  •TRIPLEX•R,W1,W2,YHR,YHI,FR,FI,GR,GI,CK,SK;
  •ARRAY•DR,DI,RD(0:N-1);
  •REAL•A1;
```

```
'COMMENT' ZUNAECHEST WERDEN DIE KOEFFIZIENTEN  $YR(J)+I*YI(J)$  DES
GEBEBENEN RECHTECKVEKTORS IN KREISSCHEIBEN
<(DR(J),DI(J)),RD(J)> EINGESCHLOSSEN. ANSCHLIESSEND WIRD DER REAL
BZW. DER IMAGINAERTEIL VON  $\exp(2*I*J*PI/N)$ ,  $J=0(1)N/2-1$ , BERECHNET
UND AUF  $YR(J)$  BZW.  $YI(J)$  ABGESPEICHERT. NACH DURCHFUEHRUNG DER FFT
WIRD DER J-TE KOEFFIZIENT DES TRANSFORMIERTEN KREISSCHEIBENVEKTORS
IN EIN RECHTECK EINGESCHLOSSEN UND REAL- BZW. IMAGINAERTEIL DIESER
RECHTECKS AUF  $YR(J)$  BZW.  $YI(J)$  ABGESPEICHERT;
```

```
'COMMENT' EINSCHLIESSUNG DER GEBEBENEN KOMPLEXEN
INTERVALLE IN KREISSCHEIBEN;
```

```
'FOR' I:=0 STEP 1 UNTIL N-1 DO
BEGIN W1:=YR(I); W2:=YI(I);
  YHR:=(POINT(INF(W1))+POINT(SUP(W1)))/2;
  YHI:=(POINT(INF(W2))+POINT(SUP(W2)))/2;
  DR(I):=MAIN(YHR); DI(I):=MAIN(YHI);
  W1:=PSP(W1); W2:=PSP(W2);
  RD(I):=SUP(SQRT(W1*W1+W2*W2)/2+PSP(YHR)+PSP(YHI));
END I;
```

```
'COMMENT' BERECHNUNG DER BENÖTIGTEN POTENZEN VON
 $\exp(2*I*PI/N)$  MIT DER RECHTECKARITHMETIK;
```

```
NH:=N/2; IG:=NH/2; IP:=IG/2;
R:=<3.1415926,3.1415926,3.1415927>/N;
YHR:=SIN(P); GI:=-4*YHR*YHR;
W1:=-GI/2; W2:=YHR*SQRT(4+GI);
FR:=YR(0); YI(IG):=<1,1,1>;
FI:=YI(0); YR(IG):=<0,0,0>;
INDA:=IG+IP; YHR:=1/SQRT(2);
YR(IP):=YI(IP); YI(INDA):=YHR;
YR(INDA):=-YHR;
'FOR' I:=1 STEP 1 UNTIL IP-1 DO
BEGIN
  K:=IG-I; JH:=IG+I; J:=NH-I;
  W1:=GI*FR+W1;
  YI(K):=YI(JH); YR(I):=FR:=FR+W1;
  YR(J):=-FR;
  W2:=GI*FI+W2;
  YI(J):=YR(K); YI(I):=FI:=FI+W2;
  YR(JH):=-FI;
END I;
```

'COMMENT' SCHNELLE FOURIERTRANSFORMATION  
MIT DER KREISSCHEIBENARITHMETIK;

```
INDIF:=NH;NGR:=1;
'FOR'IG:=1'STEP'1'UNTIL'P'DO'
'BEGIN'INDA:=0;
  'FOR'IGR:=1'STEP'1'UNTIL'NGR'DO'
    'BEGIN'IP:=0;
      'FOR'I:=0'STEP'1'UNTIL'INDIF-1'DO'
        'BEGIN'I1:=INDA+I;I2:=I1+INDIF;
          GR:=POINT(DR(I1));GI:=POINT(DI(I1));
          FR:=POINT(DR(I2));FI:=POINT(DI(I2));
          CK:=POINT(RD(I1))+POINT(RD(I2));
          YHR:=GR+FR;YHI:=GI+FI;
          DR(I1):=MAIN(YHR);DI(I1):=MAIN(YHI);
          RD(I1):=SUP(CK+PSP(YHR)+PSP(YHI));
          W1:=YR(IP);W2:=YI(IP);
          YHR:=GR-FR;YHI:=GI-FI;
          GR:=YHR*W1+YHI*W2;
          GI:=YHI*W1-YHR*W2;
          DR(I2):=MAIN(GR);DI(I2):=MAIN(GI);
          RD(I2):=SUP(CK+PSP(GR)+PSP(GI));
          IP:=IP+NGR;
        'END'I;
      INDA:=INDA+INDIF+INDIF;
    'END'IGR;
  INDIF:=INDIF//2;NGR:=NGR+NGR;
'END'IG;
```

'COMMENT' EINSCHLIESSUNG DER TRANSFORMIERTEN KREISSCHEIBEN  
IN RECHTECKE;

```
'FOR'I:=0'STEP'1'UNTIL'N-1'DO'
'BEGIN'W1:=POINT(RD(I));
  W2:=POINT(DI(I));
  A1:=INF(W2-W1);
  YR(I):=COMPOSE(A1,A1,SUP(W2+W1));
  W2:=POINT(DI(I));
  A1:=INF(W2-W1);
  YI(I):=COMPOSE(A1,A1,SUP(W2+W1));
'END'I;
```

```

'COMMENT' UMORDNEN DER Y-WERTE, SOG. BITINVERSION;
'FOR' J:=0 'STEP' 1 'UNTIL' N-1 'DO'
'BEGIN' JH:=J; K:=0; KP:=NH;
  'FOR' I:=1 'STEP' 1 'UNTIL' P 'DO'
    'BEGIN' K:=K+(JH-(JH//2)*2)*KP;
      JH:=JH//2; KP:=KP//2;
    'END' I;
  'IF' K 'GREATER' J 'THEN'
    'BEGIN' YHR:=YR(J); YHI:=YI(J);
      YR(J):=YR(K); YI(J):=YI(K);
      YR(K):=YHR; YI(K):=YHI;
    'END';
  'END' J;

'IF' SCHAL 'THEN' 'GOTO' M1;

'COMMENT' BERECHNUNG DER KOEFFIZIENTEN DES TRIGONOMETRISCHEN
INTERPOLATIONSPOLYNOMS AUS DEN KOEFFIZIENTEN DER
FOURIERTRANSFORMIERTEN;

YR(N):=YR(0); YI(N):=YI(0);
NGR:=N*N; IG:=NH//2;
YHR:=SIN(R/2); R:=-4*YHR*YHR;
W1:=-R/2; W2:=YHR*SQRT(4+R);
CK:=<1,1,1>; SK:=<0,0,0>;
'FOR' J:=0 'STEP' 1 'UNTIL' IG 'DO'
'BEGIN' FR:=CK; FI:=SK;
'FOR' I:=J, NH-J 'DO'
  'BEGIN' NI:=N-I;
    GR:=YR(I)-YR(NI); GI:=YI(I)+YI(NI);
    YHR:=GR*FR+GI*FI; YHI:=GI*FR-GR*FI;
    GR:=YR(I)+YR(NI); GI:=YI(I)-YI(NI);
    YR(I):=(YHI+GR)/NGR; YR(NI):=(GR-YHI)/NGR;
    YI(I):=(YHR-GI)/NGR; YI(NI):=(YHR+GI)/NGR;
    FR:=SK; FI:=CK;
  'END' I;
  W1:=R*CK+W1; CK:=CK+W1;
  W2:=R*SK+W2; SK:=SK+W2;
'END' J;
YI(0):=<0,0,0>; YI(N):=<0,0,0>;

M1: 'END' IFFT;

```

## 2. Die Prozedur AXB

Zu vorgegebenen Intervallen  $X(J)$ ,  $Y(J)$ ,  $J=1(1)N$ , bestimmt die Prozedur AXB eine Einschließung für die Menge  $P$  aller Parameterpaare  $(\alpha, \beta)$ , derart daß  $\alpha \leq \xi_J^B \in Y(J)$ , für mindestens ein  $\xi_J \in X(J)$ ,  $J=1(1)N$ , gilt. Im Falle, daß sämtliche Intervalle  $X(J)$  Punktintervalle sind, liefert die Prozedur (bis auf Überschätzung infolge Rundung) die optimale Intervalleinschließung von  $P$ .

### Eingabeparameter

| Name | Typ                  | Bedeutung                                                                               |
|------|----------------------|-----------------------------------------------------------------------------------------|
| N    | <u>integer</u>       | Anzahl der Intervalle $X(J)$ bzw. $Y(J)$<br>( $\hat{=}$ Dimension der Vektoren $X, Y$ ) |
| X    | <u>triplex array</u> | Vektor, dessen Komponenten die Stützstellen einschließen,                               |
| Y    | <u>triplex array</u> | Vektor, dessen Komponenten, die Stützwerte enthalten;                                   |

### Ausgabeparameter

| Name  | Typ            | Bedeutung                                                                        |
|-------|----------------|----------------------------------------------------------------------------------|
| A     | <u>triplex</u> | Einschließung für den Parameter $\alpha$                                         |
| B     | <u>triplex</u> | Einschließung für den Parameter $\beta$                                          |
| PLEER | <u>label</u>   | eine Marke, zu der gesprungen wird, sollte sich herausstellen, daß $P$ leer ist. |

```

'PROCEDURE'AXB(N)DATA:(X,Y)RESULT:(A,B),EXIT:(PLEER);

'COMMENT' DAS UNTERPROGRAMM BERECHNET EINE EINSCHLIESSUNG FUEER DIE
LOESUNGSMENGE DES INTERVALLINTERPOLATIONSPROBLEMS MIT DEN
FUNKTIONEN A*EXP(B*LN(X));

'VALUE'N;'INTEGER'N;'TRIPLEX'A,B;'TRIPLEX''ARRAY'X,Y;
'LABEL'PLEER;

'COMMENT' BERECHNUNG VON LN(A) UND B;

'BEGIN' 'INTEGER'I,J;'TRIPLEX'HILF,Z;
'FOR'I:=1'STEP'1'UNTIL'N'DO'
'BEGIN'
  X(I):=LN(X(I));Y(I):=LN(Y(I));
'END'I;
HILF:=1/(X(N)-X(N-1));
A:=(Y(N-1)*X(N)-Y(N)*X(N-1))*HILF;
B:=(Y(N)-Y(N-1))*HILF;
'IF'N'EQUAL'2'THEN'GOTO'M1;
'FOR'I:=1'STEP'1'UNTIL'N-2'DO'
'BEGIN'
  'FOR'J:=I+1'STEP'1'UNTIL'N'DO'
  'BEGIN'
    HILF:=1/(X(J)-X(I));
    Z:=(Y(I)*X(J)-Y(J)*X(I))*HILF;
    'IF'DISJUNCT(A,Z)'THEN'GOTO'PLEER;
    A:=INTSCT(A,Z);
    Z:=(Y(J)-Y(I))*HILF;
    'IF'DISJUNCT(B,Z)'THEN'GOTO'PLEER;
    B:=INTSCT(B,Z);
  'END'J;
'END'I;

'COMMENT' VERBESSERUNG DER INTERVALLE Y(I);

M1: WRITE(<<'DIE INTERVALLE Y(I) VERBESSERN SICH ZU:'',A1>>);
'FOR'I:=1'STEP'1'UNTIL'N'DO'
'BEGIN'
  HILF:=INTSCT(Y(I),A*B*X(I));
  WRITE(<<'IS,3P18.8,A1>>,I,EXP(HILF));
'END'I;
A:=EXP(A);
'END'AXB;

```

Literaturverzeichnis

- [1] ALEFELD, G.: Über die Durchführbarkeit des Gaußschen Algorithmus bei Gleichungen mit Intervallen als Koeffizienten,  
Computing Suppl. 1, 15-19 (1977)
- [2] ALEFELD, G.; HERZBERGER, J.: Einführung in die Intervallrechnung,  
Bibliographisches Institut, Mannheim/Wien/Zürich (1974)
- [3] BACHMANN, K.-H.: Einschließung der Nullstellen von Intervallpolynomen,  
Beitr. Numer. Math. 4, 23-32 (1975)
- [4] BARTH, W.; NUDING, E.: Optimale Lösung von Intervallgleichungssystemen,  
Computing 12, 117-125 (1974)
- [5] BEECK, H.: Zur scharfen Außenabschätzung der Lösungsmenge bei linearen Intervallgleichungssystemen,  
Z. Angw. Math. Mech. 54, T208-T209 (1974)
- [6] BELLMANN, R.: Methods of nonlinear analysis Vol. I,  
Academic Press, New York/London (1970)
- [7] BERTALANFFY, L. von: Principles and theory of growth,  
in:  
NOWINSKI, W. W.: Fundamental aspects of normal and malignant growth,  
Elsevier Publ. Comp., Amsterdam/London/New York (1960)
- [8] BHATTACHARJEE, G. P.; MAJUMDER, K. L.: On the computation of Hermite interval interpolating polynomials,  
Computing 19, 73-83 (1977)

- [9] BHATTACHARJEE, G.P.; MAJUMDER, K.L.: Multivariate interval interpolation,  
J.Comput.Appl.Math. 4, 295-300 (1978)
- [10] BÖHMER, K.: Spline Funktionen,  
B.G.Teubner Verlag, Stuttgart (1974)
- [11] BOOR, C. de: A practical guide to splines,  
Springer Verlag, New York/Heidelberg/Berlin (1978)
- [12] BOOR, C.de; Pinkus, A.: Backward error analysis for totally positive linear systems,  
Numer.Math. 27, 485-490 (1977)
- [13] BRIGHAM, O.E.: The fast Fourier transform,  
Prentice-Hall, Englewood Cliffs, N.J. (1974)
- [14] BRUTMAN, L.: On the Lebesgue function for polynomial interpolation,  
SIAM J.Numer.Anal. 15, 694-704 (1978)
- [15] COOLEY, J.W.; TUCKEY, J.W.: An algorithm for machine calculation of complex Fourier series,  
Math.Comput. 19, 297-301 (1965)
- [16] CRYER, C.W.: Some properties of totally positive matrices,  
Linear Algebra and Appl. 15, 1-25 (1976)
- [17] DOBRONETS, B.S.; SHOKIN, Yu.I.; YULDASHEV, Z.Kh.:  
Zadacha interpolatsii intervalnogo analiza,  
Voprosy Vyčisl. i. Prikl. Mat. 31, 9-16 (1975)
- [18] EHLICH, H.; ZELLER, K.: Auswertung der Normen von Interpolationsoperatoren,  
Math. Ann. 164, 105-112 (1966)

- [19] GAIER,D.; HÜBNER,O.: Schnelle Auswertung von  $Ax$  bei Matrizen zyklischer Bauart, Toeplitz- und Hankel-Matrizen,  
Mitt.Math.Sem. Giessen 121, 27-38 (1976)
- [20] GANTMACHER,F.R.; KREIN,M.G.: Oszillationsmatrizen, Oszillationskerne und kleine Schwingungen mechanischer Systeme,  
Akademie-Verlag, Berlin (1960)
- [21] GARLOFF,J.: Intervallinterpolation mit der Funktionenklasse  $ax^b$  ,  
Freiburger Intervall-Berichte 78/1,  
Inst. f. Angew. Math., Universität Freiburg i.Br. (1978)
- [22] GARLOFF,J.: Optimale Schranken bei Intervallinterpolation mit Polynomen und mit Funktionen  $ax^b$  ,  
Z. Angew. Math. Mech. 59, T59-T60 (1979)
- [23] GARLOFF,J.: Über trigonometrische Intervallinterpolation,  
Freiburger Intervall-Berichte 79/5,  
Inst. f. Angew. Math., Universität Freiburg i.Br.,  
31-103 (1979)
- [24] GARLOFF,J: Zur intervallmäßigen Durchführung der schnellen Fouriertransformation,  
Kurzvortrag, gehalten auf der Jahrestagung der GAMM in Wiesbaden 1979  
eingereicht bei der Z. Angew. Math. Mech.
- [25] GENTLEMAN,W.M.; SANDE,G.: Fast Fourier transforms for fun and profit,  
1966 Fall Joint Computer Conference, AFIPS Proc. 29, Washington D.C., Spartan Books,  
563-578 (1966)

- [26] HENRICI, P.: Applied and computational complex analysis  
Vol. I,  
John Wiley & Sons, New-York/London/Sydney (1974)
- [27] HENRICI, P.: Fast Fourier methods in computational complex  
analysis,  
(Von Neumann Lecture, SIAM National Meeting 1978)  
Research Report No. 78-04  
Seminar für Angewandte Mathematik, ETH Zürich (1978)
- [28] HERZBERGER, J.: Note on a bounding technique for  
polynomial functions,  
SIAM J. Appl. Math. 34, 685-686 (1978)
- [29] KARLIN, S.: Total positivity vol. I,  
Stanford University Press, Stanford, Calif. (1968)
- [30] KRAWCZYK, R.: Approximation durch Intervallfunktionen,  
Interner Bericht des Inst. f. Informatik 69/7,  
Universität Karlsruhe (1969)
- [31] KUTTLER, J.R.: A fourth-order finite-difference approxima-  
tion for the fixed membrane eigenproblem,  
Math. Comput. 25, 237-256 (1971)
- [32] LANCZOS, C.: Applied analysis,  
Prentice Hall, New York (1956)
- [33] MAJUMDER, K.L.; BHATTACHARJEE, G.P.: Some algorithms for  
interval interpolating polynomials,  
Computing 16, 305-317 (1976)
- [34] METELMANN, K.: Inverspositive Bandmatrizen und totalnicht-  
negative Green'sche Matrizen,  
Dissertation, Universität Köln (1972)

- [35] METELMANN, K.: Ein Kriterium für den Nachweis der Total-nichtnegativität von Bandmatrizen,  
Linear Algebra and Appl. 2, 163-171 (1973)
- [36] MILLER, W.: Computational complexity and numerical stability,  
SIAM J. Comput. 4, 97-107 (1975)
- [37] MOORE, R.E.: Interval analysis,  
Prentice Hall, Inc., Englewood Cliffs, N.J. (1966)
- [38] NICKEL, K.: Die Überschätzung des Wertebereichs einer Funktion in der Intervallrechnung mit Anwendungen auf lineare Gleichungssysteme,  
Computing 18, 15-36 (1977)
- [39] NICKEL, K.: Intervall-Mathematik,  
Z. Angew. Math. Mech. 58, T72-T85 (1978)
- [40] NUDING, E.: Intervallarithmetik, Teil 2, Skriptum zur Vorlesung,  
Rechenzentrum der Universität Heidelberg (1972)
- [41] PRONY, R.: Essai experimental et analytique sur les lois de la dilatabilité des fluides élastiques et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alkool, à différentes températures,  
Journal de l'Ecole Polytechnique 1, Heft 2, 24-76 (1795)
- [42] RAMOS, G.U.: Roundoff error analysis of the fast Fourier-transform,  
Math. Comput. 25, 757-768 (1971)
- [43] RATSCHEK, H.: Ergebnisse einer Untersuchung über die Struktur von Intervallpolynomen,  
Berichte der Gesellschaft für Mathematik und Datenverarbeitung Bonn, Nr. 52 (1972)

- [44] REICHMANN, K.: Abbruch beim Intervall-Gauß-Algorithmus,  
Computing 22, 355-361 (1979)
- [45] ROKNE, J.: Explicit calculation of the Lagrangian interval interpolating polynomial,  
Computing 2, 149-157 (1972)
- [46] RUTISHAUSER, H.: Der Quotienten-Differenzen-Algorithmus,  
Birkhäuser Verlag, Basel (1957)
- [47] SAUER, R. und SZABÓ, I.: Mathematische Hilfsmittel des Ingenieurs Teil III,  
Springer Verlag, Berlin/Heidelberg/New York (1968)
- [48] SCHAEFER, H.H.: Topological vector spaces,  
3. Auflage, Springer Verlag, Berlin/Heidelberg/  
New York (1971)
- [49] SCHWARZ, H.R.: Elementare Darstellung der schnellen Fourier-Transformation,  
Computing 18, 107-116 (1977)
- [50] STOER, J.: Einführung in die Numerische Mathematik I,  
2. Auflage, Springer Verlag, Berlin/  
Heidelberg/ New York (1976)
- [51] VARGA, R.S.: Matrix iterative analysis,  
Prentice Hall, Inc., Englewood Cliffs, N.J.  
(1962)
- [52] VARGA, R.S.: On recurring theorems on diagonal dominance,  
Linear Algebra and Appl. 13, 1-9 (1976)

- [53] WERNER, H.; SCHABACK, R.: Praktische Mathematik II,  
Springer Verlag, Berlin/Heidelberg/  
New York (1972)
- [54] WHITNEY, A.M.: A reduction theorem for totally  
positive matrices,  
J. Analyse Math. 2, 88-92 (1952)
- [55] WIPPERMANN, H.-W. (Hrsg.):  
APOSTOLATOS, N.; KRAWCZYK, R.; KULISCH, U.;  
LORTZ, B.; NICKEL, K.; WIPPERMANN, H.-W.  
The algorithmic language Triplex-ALGOL 60,  
Numer. Math. 11, 175-180 (1968).

## Lebenslauf

Name: Jürgen Garloff

Geboren: 23. 3. 1950 in Köln-Lindenthal

Eltern: Dipl.-Kaufmann Dr. rer. pol. Fritz Garloff und  
Fachärztin Dr. med. Liselotte Garloff, geb. Wunderlich

Schule: 1956 - 1960 Grundschule in Bonn-Rad Godesberg  
und in Köln-Buchforst  
1960 - 1968 Gymnasium in Köln-Mülheim und  
in Hinterzarten (Südschwarzwald)

Studium: 1968 - 1972 Studium der Soziologie, Philosophie,  
Geschichte und Mathematik (mit Nebenfach  
Wirtschaftswissenschaften) an  
der Universität Köln  
1972 - 1976 Studium der Mathematik (mit Nebenfach  
Wirtschaftswissenschaften) an der  
Universität Heidelberg  
Mai 1976 Diplom in Mathematik

Beruf: Seit Oktober 1976 wissenschaftlicher Angestellter  
am Institut für Angewandte Mathematik der Universität Freiburg i.Br.

# F r e i b u r g e r

## I n t e r v a l l - B e r i c h t e

1978

- 1 J.Garloff: Intervallinterpolation mit der  
Funktionenklasse  $ax^b$
- 2 N.C.Börsken: Komplexe Kreis-Standardfunktionen
- 3 J.M.Yohe: A Semi-portable Interval Arithmetic  
Package for FORTRAN
- 4 S.M.Markov: Extended Interval Arithmetic and some  
Applications (Vortrag auf der Intervall-Tagung Seefeld 1978)
- 5 F.Bierbaum: Die Numerische Konvergenz von Programmen  
(Dr.-Dissertation)
- 6 K.L.Hellmig: Existenz von Lösungen bei Anfangswertproblemen  
von Intervall-Differential-Gleichungen (Diplomarbeit)
- 7 K.Nickel: Was ist eigentlich Angewandte Mathematik ?  
(Antrittsvorlesung an der Universität Freiburg i.Br.)
- 8 D.Klaua: Mengen-Intervalle
- 9 J.Rohn: Input-Output Planning With Inexact Data
- 10 K.Nickel: Die Konstruktion von a-priori-Schranken zur  
Lösung eines Randwertproblems mit finiten Elementen,  
Teil I

F r e i b u r g e r  
I n t e r v a l l - B e r i c h t e

1979

- 1 J.Garloff, K.-P.Schwierz: Intervall-Mathematik.  
Eine Literaturübersicht. Nachtrag
- 2 W.J.Beiser: Intervall-Newton-Verfahren zur Bestimmung von  
Nullstellen reeller Funktionen einer Veränderlichen
- 3 M.J.Schneider: Die Definition der Sprache "Langzahl-  
Triplex-ALGOL 60" und ihre Implementierung durch einen  
Praecompiler. (Diplomarbeit)
- 4 K.Nickel: Schranken für die Lösungsmengen von Funktional-  
Differentialgleichungen
- 5 K.Reichmann: Abbruch beim Intervall-Gauss-Algorithmus  
J.Garloff: Über trigonometrische Intervallinterpolation
- 6 W.Solak: Iterative Methods for Convex and Inverse Monotonic  
Operators  
M.Raith: Existenz und Eindeutigkeit von zentrierten Kugel-  
Erweiterungen
- 7 B.Bücher: Einführung in die Intervallmathematik im Rahmen  
einer Arbeitsgemeinschaft am Gymnasium. (Diplomarbeit)
- 8 E.Gardeñes, A.Trepat: The Interval Computing System  
SIGLA-PL/1(0)
- 9 K.Nickel: Ein global konvergentes Kugel-Newton-Verfahren
- 10 S.Markov, N.Dimitrova: Rechengesetze der erweiterten Intervall  
arithmetik  
F.Wildenauer: Domains with all Solutions of non-linear  
Problems with non-inverse-isotone Operators

F r e i b u r g e r  
I n t e r v a l l - B e r i c h t e

1980

- 1 J.Conradt: Ein Intervallverfahren zur Einschließung des Fehlers einer Näherungslösung bei Anfangswertaufgaben für Systeme von, gewöhnlichen Differentialgleichungen. (Diplomarbeit)
- 2 . Interval Symposium 1980. Abstracts
- 3 R.Krawczyk: Zur Konvergenz iterierter Mengen.  
D.Klausa: Intervall-Komponenten der nichtarchimedischen Zahlengeraden.
- 4 K.Reichmann: Die Konvergenz von Intervall-(Potenz-)Reihen mit Anwendungen auf Intervall-Anfangswertprobleme.  
(Dr. Dissertation)
- 5 J.Garloff: Untersuchungen zur Intervallinterpolation.  
(Dr. Dissertation)





